Marzo 1989
Numero 247
Anno XXII
Volume XLII

I terremoti profondi

Da quando se ne è accertata l'esistenza, 60 anni fa, si pone uno stimolante interrogativo: come è possibile la fratturazione di rocce alle temperature e pressioni che si riscontrano a centinaia di chilometri di profondità?

di Cliff Frohlich

Nella maggior parte dei terremoti la crosta terrestre si frattura come porcellana. Le tensioni meccaniche si accumulano fino a originare fratture crostali ad alcuni chilometri di profondità e lo scorrimento di blocchi che ne consegue scarica la tensione stessa. Alcuni terremoti, tuttavia, hanno origine dove non è possibile la fratturazione, cioè nel mantello terrestre, a centinaia di chilometri di profondità. Si suppone che le altissime pressioni presenti nel mantello impediscano alle rocce di fratturarsi, anche quando le sollecitazioni compressive sono di tale entità da deformare le rocce stesse quasi fossero stucco. Come è dunque possibile che avvengano terremoti a tali profondità?

Questi misteriosi eventi profondi sono abbastanza comuni. Dal 1964 l'International Seismological Centre (ISC) di Londra ha catalogato più di 60 000 terremoti a profondità maggiori di 70 chilometri, pari al 22 per cento di tutti i terremoti localizzati. Di essi, alcuni hanno avuto effetti disastrosi. Sebbene quasi tutti i terremoti catastrofici abbiano origine in superficie, cioè a una profondità minore di 50 chilometri, l'evento sismico che il 4 marzo 1977 devastò Bucarest aveva l'ipocentro a 100 chilometri di profondità.

Lo studio degli eventi profondi ha contribuito a fare avanzare di molto le conoscenze in campo geofisico. La distribuzione geografica di questo tipo di terremoti va a ulteriore conferma di quella «grande teoria unificatrice» della geofisica moderna che è la tettonica a zolle. Questi eventi hanno rappresentato inoltre ideali sorgenti di energia per gli studi della parte più interna della Terra, studi che tentano di risalire alla struttura interna attraverso il comportamento delle onde sismiche che si propagano nei vari strati del pianeta. Ora gli stessi terremoti profondi potrebbero svelarci i loro segreti: le osservazioni sismiche, unitamente agli esperimenti di laboratorio sul comportamento delle rocce ad alte pressioni hanno condotto a spiegazioni plausibili su come questi eventi abbiano luogo.

La profondità a cui i terremoti hanno origine fu oggetto di controversia per decenni fino al 1927, quando il sismologo giapponese Kiyoo Wadati dimostrò in modo lampante l'esistenza di eventi sismici profondi. Notando che, in alcuni casi, il movimento sismico intenso era circoscritto a zone ristrette, alcuni ricercatori avevano supposto che il centro di liberazione dell'energia (fuoco o ipocentro del sisma) si trovasse a non più di qualche chilometro dalla superficie. Altri erano giunti a ipotizzare profondità focali (ovvero profondità degli ipocentri) addirittura di 1200 chilometri, tentando di determinare la provenienza delle onde sismiche sulla base delle registrazioni dei sismografi - ancora alquanto rudimentali - di cui disponevano.

La controversia si acuì con l'ampliamento delle conoscenze sulle onde sismiche e sulle loro modalità di propagazione. I ricercatori impararono a riconoscere sui sismogrammi diversi tipi di onde interne (quelle che si propagano *attraverso* la Terra, diverse quindi dalle onde superficiali, che seguono la superficie terrestre). Le prime ad arrivare sono le onde *P*, o primarie (dette anche onde di compressione, o longitudinali) che, con la stessa modalità delle onde acustiche, si propagano alternando zone di alta e bassa pressione. Esse sono seguite dalle onde *S* (secondarie, o trasversali) in cui l'oscillazione è ortogonale alla direzione di propagazione. Confrontando i tempi di arrivo a diverse stazioni delle onde *P* e *S* di un dato evento sismico si poté osservare che i tempi di percorrenza dipendevano sia dalla distanza dall'ipocentro, sia dalla struttura interna della Terra.

Dato quindi un modello di struttura interna della Terra e dati i tempi di arrivo delle onde sismiche in un certo numero di stazioni, si potrebbe in teoria localizzare l'ipocentro dell'evento. Nel 1922 H. H. Turner, direttore del centro di raccolta e interpretazione di dati sismologici che in seguito sarebbe diventato l'ISC, applicò questo metodo in una pubblicazione stimolante e controversa. Sulla base dei dati di stazioni distribuite in tutto il mondo, Turner propose di classificare i terremoti in tre livelli di profondità. Gli eventi «ad alto fuoco» hanno il loro ipocentro in prossimità della superficie, ma i terremoti normali, ovvero quelli del tipo più comune, avvengono fino a profondità di 150 chilometri. Gli eventi «a fuoco profondo» si originano fino a 650 chilometri di profondità.

L'approccio di Turner era buono, ma

Un terremoto profondo dalle conseguenze catastrofiche è un evento raro. L'unico che ha prodotto gravi danni negli ultimi decenni è avvenuto a Bucarest il 4 marzo 1977, a una profondità di 100 chilometri. I pilastri al pianterreno di questo edificio hanno ceduto, facendo sprofondare di un piano la facciata d'angolo. Le vittime in città sono state 1500. La fotografia è stata ripresa da Neculai Mandrescu, dell'Istituto di fisica terrestre di Bucarest.

i dati di cui egli disponeva erano approssimativi rispetto agli standard attuali: la conoscenza che si aveva a quel tempo sulla struttura profonda della Terra era lacunosa e gli orologi imprecisi e mal sincronizzabili, così da falsare i tempi di arrivo delle onde sismiche di secondi o addirittura di minuti. Gli argomenti addotti da Turner convinsero ben pochi dei suoi contemporanei. S.K. Banerji dell'Osservatorio di Bombay fece notare che, se il grosso dei terremoti avviene a profondità di 150 chilometri o più, pochi dovrebbero essere gli eventi in grado di produrre forti onde superficiali, e tuttavia proprio il catalogo di Turner riportava onde superficiali in abbondanza. Harold Jeffreys dell'Università di Cambridge avanzò un'obiezione più fondamentale: sostenne che i terremoti non potevano semplicemente avere luogo a simili profondità.

Al di sotto di profondità di circa 50 chilometri - sosteneva Jeffreys - la temperatura e la pressione rendono plastico il materiale che costituisce il mantello, di per sé fragile e fratturabile. A sostegno della sua argomentazione Jeffreys rimarcò il fatto che, dal termine dell'ultima glaciazione, il Canada e l'Europa settentrionale, liberati dal peso delle coltri glaciali, hanno subito un sollevamento, così da far pensare che il mantello sottostante sia capace di scorrere plasticamente sotto sollecitazione, come un liquido estremamente viscoso. Egli citò inoltre esperienze di laboratorio che confermavano come, ad alte temperature e pressioni, le rocce si deformino gradualmente anziché fratturarsi.

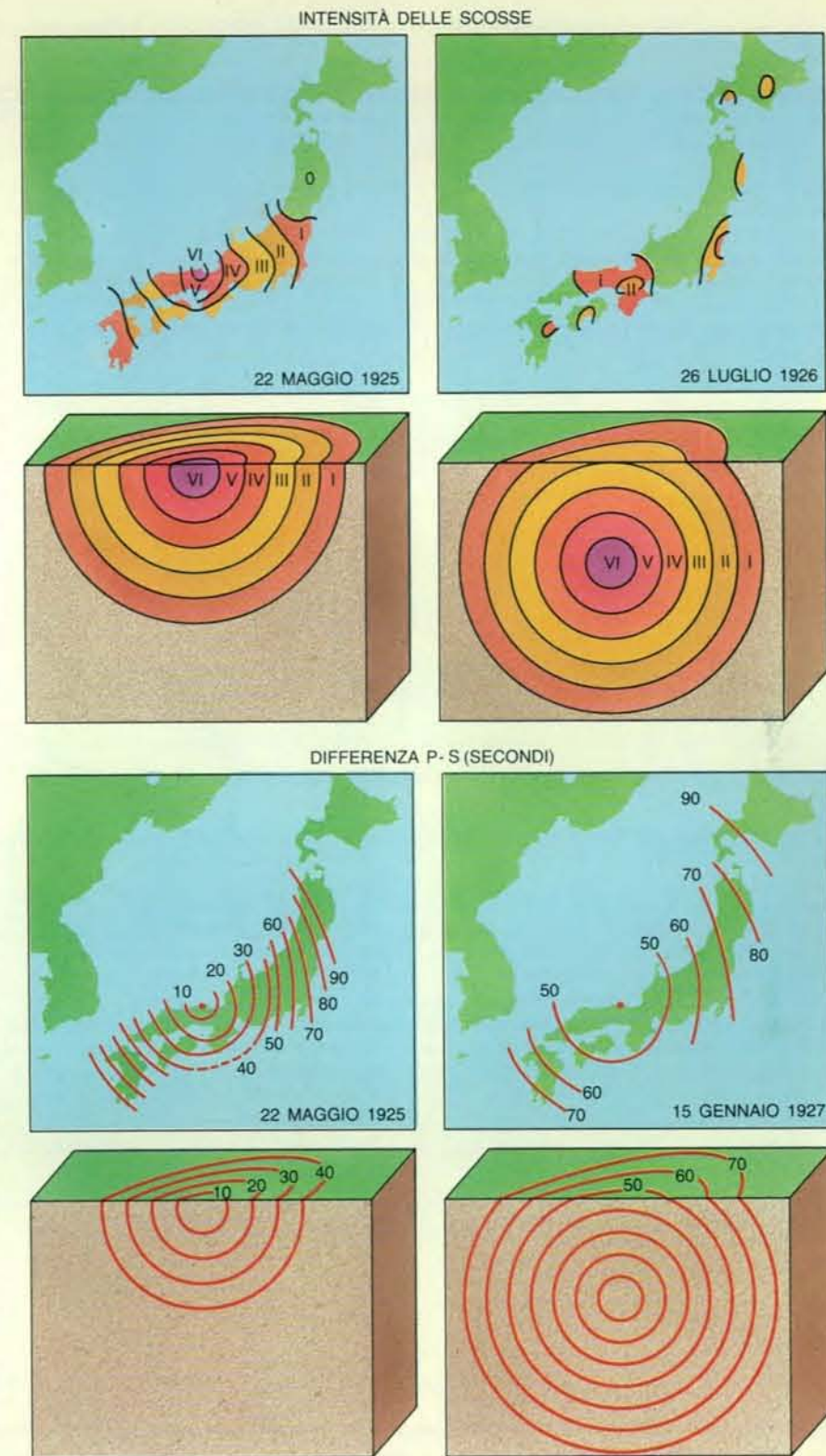
Kiyoo Wadati, impiegato venticinquenne del Servizio meteorologico giapponese, non confutò le argomentazioni di Jeffreys, ma si limitò a presentare prove evidenti dell'esistenza di terremoti molto profondi. La frequenza e la gravità dei terremoti in Giappone avevano indotto il Governo giapponese a istituire quella che all'epoca era la miglior rete di stazioni sismografiche al mondo. Quindi Wadati disponeva di dati abbondanti e poté applicare nuovi metodi per la determinazione della profondità dei terremoti. Anziché confrontare i tempi assoluti di arrivo a diverse stazioni sismografiche, come aveva fatto Turner, egli si basò su una differenza di tempo che poteva essere facilmente misurata in ogni stazione anche con orologi imprecisi: l'intervallo tra l'arrivo delle onde *P* e quello delle onde *S*, più lente. Poiché ogni tipo di onda si propaga a velocità abbastanza costante, l'intervallo cresce proporzionalmente alla distanza della stazione dall'ipocentro del terremoto.

Wadati scoprì che per la maggior parte dei terremoti l'intervallo era piuttosto piccolo in prossimità dell'epicentro, ovvero del punto di più intensa scossa. In qualche caso però il ritardo era lungo anche all'epicentro. Wadati trovò un'analogia a questo fatto analizzando i dati sull'intensità delle scosse. Per la maggior parte dei terremoti le scosse più intense erano limitate a un'area ristretta e si affievolivano rapidamente allontanandosi dall'epicentro, ma per altri il picco di intensità era meno accentuato e diffuso su un'area più ampia. Sia gli intervalli *P-S*, sia le diverse distribuzioni di intensità in superficie facevano pensare all'esistenza di due tipi di terremoti: eventi superficiali, nei quali l'ipocentro si trova appena al di sotto dell'epicentro, ed eventi profondi, con ipocentro a centinaia di chilometri sotto la superficie.

Altri ricercatori applicarono le tecniche di Wadati ai dati sismici in altre aree geografiche, confermando i suoi risultati: i terremoti «normali» avevano una profondità focale di 50 chilometri o meno, ma alcuni eventi avevano un'origine molto più profonda, fino a 600 chilometri e anche più. Turner si era sbagliato sulla profondità dei terremoti normali, ma gli eventi profondi esistevano effettivamente. Anche Banerji aveva avuto ragione: i sismogrammi degli eventi riconosciuti come profondi mostravano solamente onde superficiali deboli.

Cosa ne era dunque dell'asserzione di Jeffreys, secondo cui le rocce del mantello a più di 50 chilometri di profondità sarebbero troppo plastiche per immagazzinare le tensioni necessarie allo scatenarsi di un terremoto? Un'osservazione di Wadati faceva presagire parte della risposta: i terremoti profondi non avvengono nella normale roccia del mantello. Nel 1935 Wadati pubblicò una mappa dei siti dei terremoti giapponesi e delle loro profondità focali. Egli aveva trovato che gli ipocentri si disponevano lungo linee approssimativamente parallele, la cui profondità aumentava in modo costante dalla costa orientale del Giappone verso ovest. Wadati commentava: «La possibilità di tracciare isolinee delle profondità ipocentrali porta a pensare che esista nella crosta qualcosa di simile a una *superficie debole*... dove è possibile che avvengano i terremoti. Questa superficie si estende immergendosi ad angolo costante nella crosta in prossimità dell'arcipelago giapponese».

La profondità dei terremoti in altre parti del mondo, continuava Wadati, definisce superfici in immersione dello stesso genere: «I terremoti profondi tendono ad aver luogo sul lato più vicino a un continente e quelli superficiali sul lato opposto, che in molti casi confina con un mare molto profondo. Questa tendenza sembra potersi osservare in molte regioni vulcaniche della Terra».

Quasi tutti i terremoti profondi si conformano allo schema descritto da Wadati. Laddove essi sono comuni - generalmente ai margini di un oceano profondo - definiscono un «piano» inclinato che si estende dalla prossimità della superficie a una profondità di 600 chilometri o più. Questi piani inclinati sono noti come zone di Wadati-Benioff da quando Wadati e il sismologo Hugo Benioff ne compila-

Le manifestazioni di superficie distinguono i terremoti profondi e quelli poco profondi. Sono confrontati i dati, raccolti dal sismologo giapponese Kiyoo Wadati, sull'intensità delle scosse e sugli intervalli tra l'arrivo delle onde *P* (di compressione) e delle onde *S* (di taglio) per alcuni terremoti avvenuti tra il 1925 e il 1927. Le scosse erano molto intense presso l'epicentro di un evento del 1925, ma diminuivano rapidamente con la distanza (*in alto a sinistra*); un sisma nel 1926 produsse scosse meno intense che decrescevano più lentamente (*in alto a destra*). La differenza *P-S* all'epicentro del primo evento era minore di 10 secondi, ma il ritardo cresceva rapidamente con la distanza (*in basso a sinistra*). La minima differenza *P-S* per un evento del 1927 fu più lunga - circa 40 secondi - ma con un aumento più lento (*in basso a destra*). Wadati concluse che il terremoto del 1925 avesse ipocentro superficiale e che i successivi lo avessero circa 400 chilometri più in profondità (*viste in spaccato*).

rono la carta negli anni quaranta e cinquanta e la loro scoperta costituì un passo decisivo verso il nuovo paradigma geofisico che doveva prendere piede negli anni sessanta.

Questa nuova visione sintetica, la teoria della tettonica a zolle, dà ragione di molte manifestazioni geologiche superficiali e spiega in gran parte la storia geologica con il movimento delle zolle crostali che ricoprono la superficie della Terra. Le zolle divergono e si allontanano in corrispondenza delle dorsali medio-oceaniche; laddove esse collidono, generalmente al margine di un bacino oceanico, provocano il sollevamento di montagne e modellano i margini continentali. Questa dinamica crostale è l'espressione in superficie dei moti di convezione che avvengono all'interno del mantello terrestre. Il materiale più caldo risale entro il mantello per poi circolare orizzontalmente in prossimità della superficie terrestre. Grosso modo i 50 chilometri più superficiali di questo flusso orizzontale si raffreddano e formano le zolle rigide, che comprendono crosta e parte del mantello sottostante.

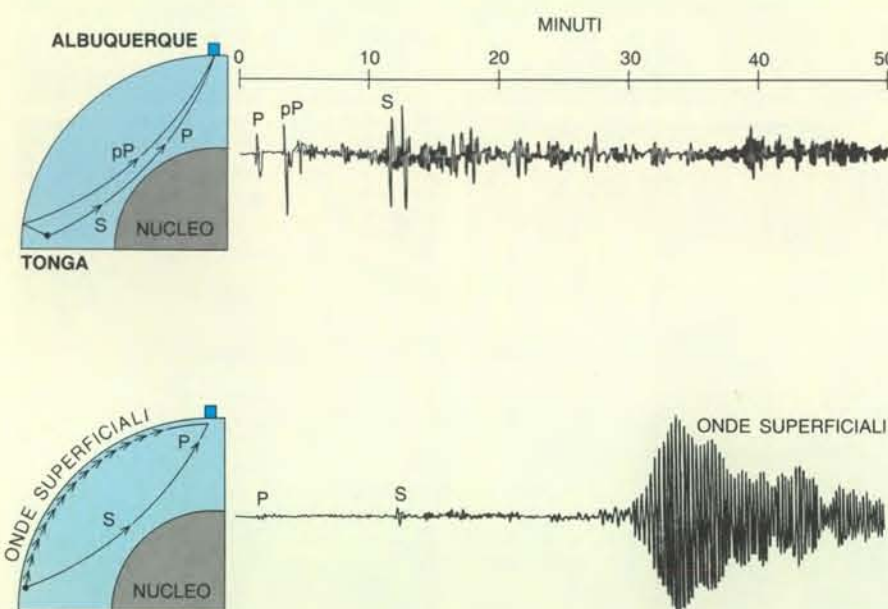
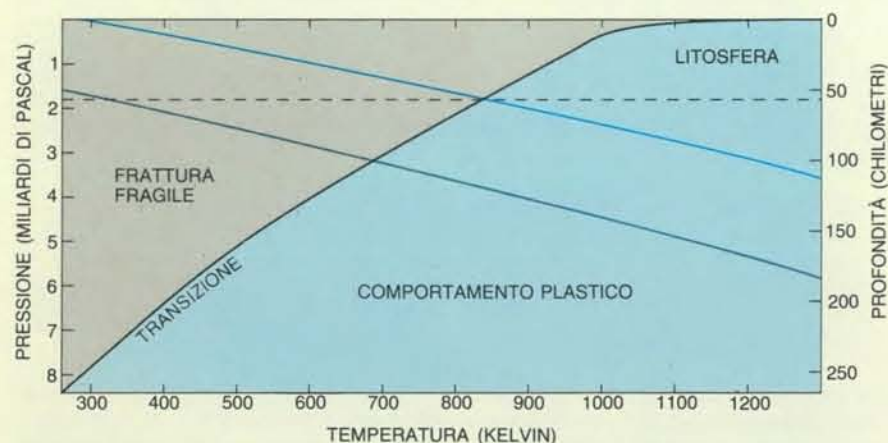
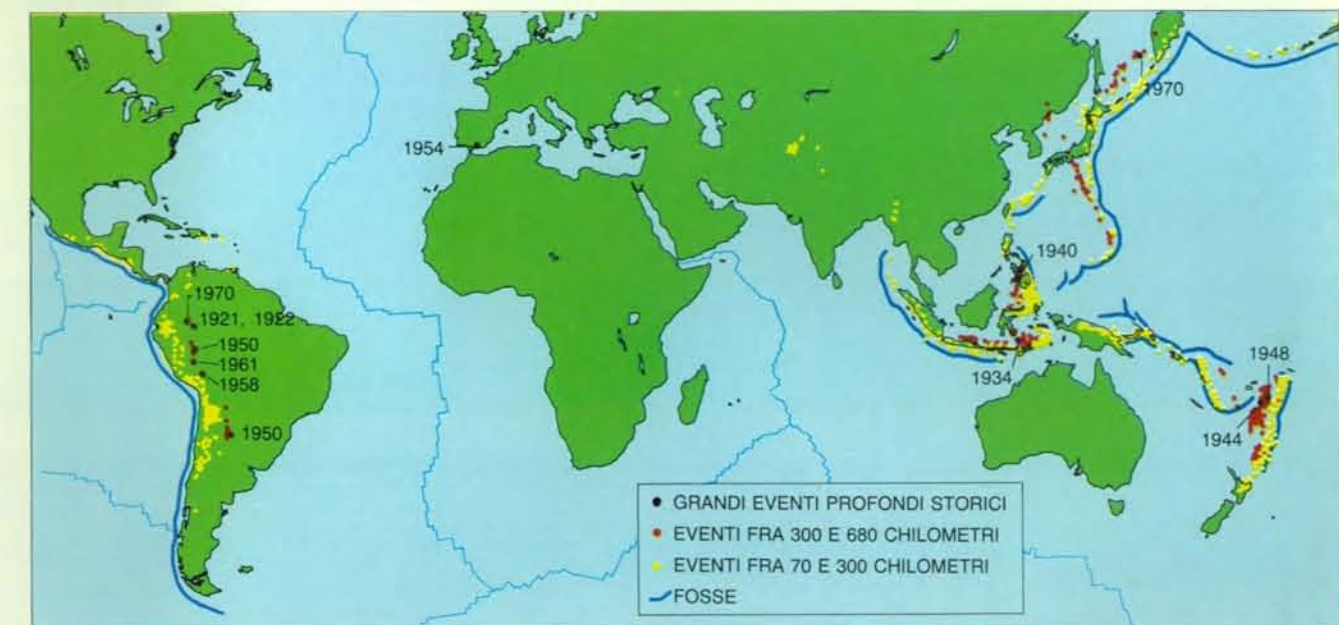
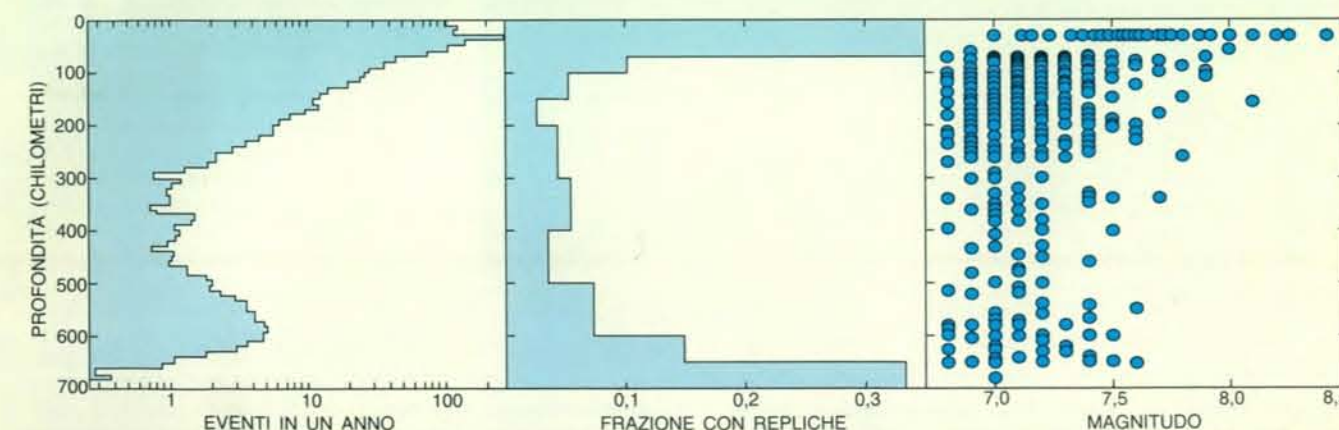
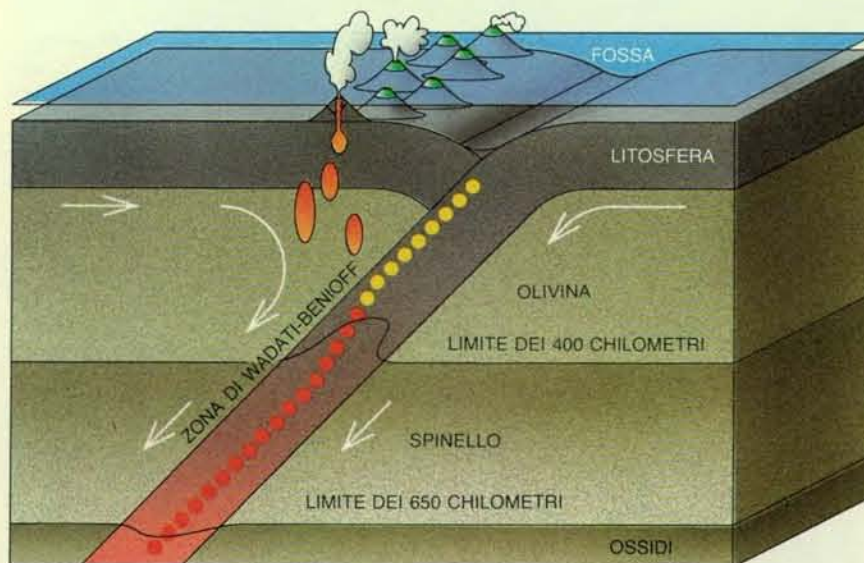
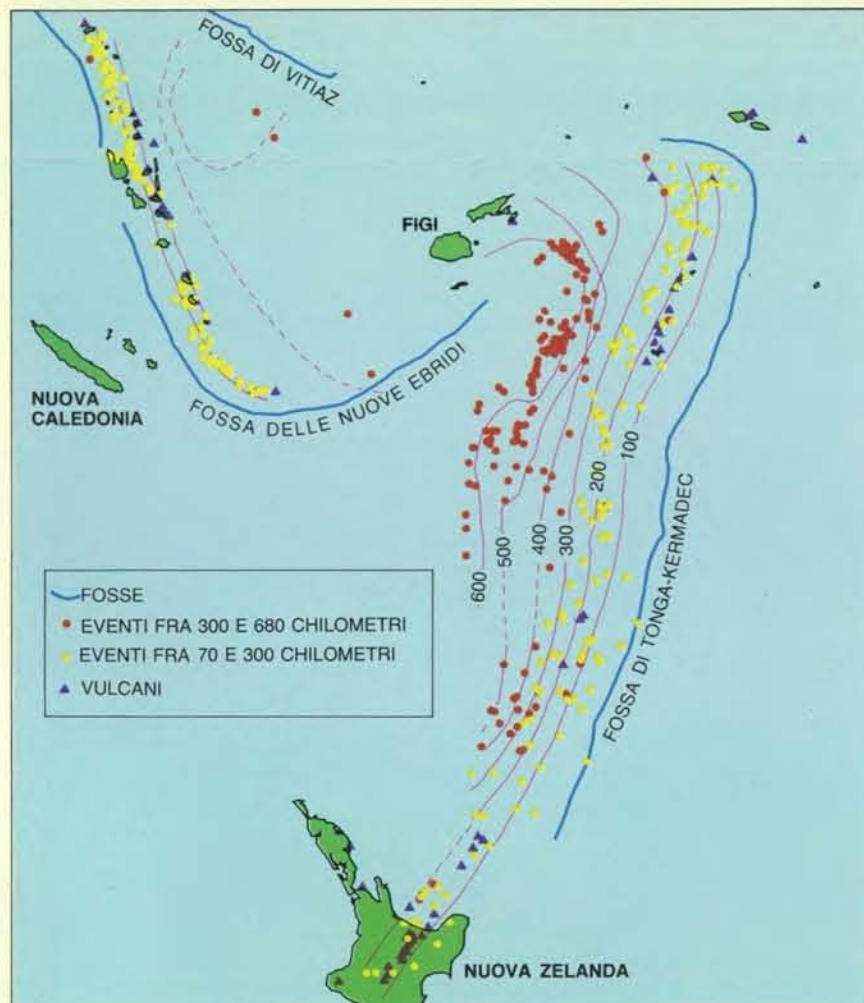
Nella cella di convezione il moto verso il basso del materiale freddo avviene laddove le zolle convergono. Qui una delle due zolle va in subduzione, ossia si flette al di sotto dell'altra zolla e affonda nel mantello. I terremoti profondi hanno contribuito a stabilire la realtà del fenomeno di subduzione, dal momento che ci si accorse che essi avvengono su una sorta di lastrone in via di immersione di cui la zona di Wadati-Benioff ricalca la forma. La profonda fossa che generalmente si riscontra appena «a mare» della zona di Wadati-Benioff - il «mare profondo» della descrizione di Wadati - segna l'inclinazione verso il basso della zolla che sta andando in subduzione; la catena di vulcani che spesso si forma in prossimità è alimentata dal materiale fuso che risale dalla zolla in subduzione. Wadati era stato profeta: nel suo articolo del 1935 aveva ipotizzato che i terremoti e i vulcani giapponesi potessero essere il risultato della deriva dei continenti (teoria precorritrice della tettonica a zolle) proposta circa 20 anni prima da Alfred Wegener.

Questo schema geofisico risponde già parzialmente all'obiezione di Jeffreys. I terremoti profondi in una zona di Wadati-Benioff avvengono in una roccia centinaia di gradi più fredda del mantello circostante e quindi meno plastica e più atta a immagazzinare energia elastica. Ma anche altri fattori sembrano essere all'origine della distribuzione dei terremoti profondi. Gli ipocentri, per esempio, non sono uniformemente distribuiti lungo la zona di Wadati-Benioff; le variazioni nella frequenza dei terremoti invece sembrano coincidere con profondità alle quali la struttura cristallina delle rocce del mantello cambia e passa a una fase più densa a causa dell'aumento di pressione.

Alla profondità di circa 400 chilometri, per esempio, gli studi sulla propagazione delle onde sismiche mostrano un improvviso aumento nella velocità di queste, che indica un aumento di densità della roccia. A questa profondità l'olivina - il silicato che costituisce, con proporzioni variabili di ferro e magnesio, il principale componente del mantello e della zolla subdotta - passa alla fase cristallina più densa dello spinello. Circa a questa profondità il numero di terremoti profondi scende a un valore minimo.

Alcune zone di subduzione non presentano più attività sismica al di sotto di questo primo livello di transizione. Le zone che invece hanno terremoti a queste profondità mostrano il loro più alto grado di attività fra questo primo livello di transizione e un secondo, più misterioso, situato a circa 650 chilometri di profondità. Anche questo livello è caratterizzato da un brusco aumento nella velocità delle onde sismiche, ma gli studiosi discutono se ciò sia dovuto a una seconda transizione di fase o a un cambiamento di composizione. Comunque l'attività sismica cade bruscamente a zero in prossimità di questo secondo livello.

Nel tentativo di determinare la massima profondità degli eventi profondi, Philip B. Stark, allora all'Università del Texas ad Austin, e io applicammo diversi metodi di analisi. Fra l'altro, esaminammo gli intervalli fra l'arrivo delle normali onde P e quello delle onde pP (onde di pressione che viaggiano verso la superficie e sono riflesse nuovamente nell'interno della Terra fino a stazioni sismografiche lontane). Trovammo che gli eventi più profondi registrati erano avvenuti tra i 680 e i 690 chilometri. Beth A. Rees ed Emile A. Okal della Northwestern University effettuarono un'analisi simile, con risultati paragonabili.

Il venir meno dell'attività sismica è troppo improvviso per essere il risultato di un graduale aumento di plasticità del lembo di crosta in subduzione che viene riscaldato dal mantello circostante. Potrebbe darsi piuttosto che il lembo non possa penetrare al di sotto del limite dei 650 chilometri. Se ciò fosse vero la convezione sarebbe confinata al mantello

Sismogrammi di terremoti superficiali e profondi avvenuti nell'arcipelago di Tonga sono stati registrati ad Albuquerque, nel New Mexico, a un quarto di globo di distanza. L'evento profondo (*in alto*) produsse intense onde *P* ed *S*, che attraversavano la Terra a velocità differenti. Alcune delle onde *P* assumevano la forma di onde *pP*: dopo essersi propagate dall'ipocentro alla superficie erano riflesse nuovamente all'interno. A causa della profondità - circa 625 chilometri - l'evento produsse solo poche onde di superficie. Sia le onde *P* sia le *S* dell'evento superficiale (*in basso*) erano relativamente deboli; la maggior parte dell'energia fu osservata sotto forma di onde superficiali, l'ultima forma di segnale ad arrivare.

L'enigma dei terremoti profondi è riassunto dal diagramma che mostra le condizioni di temperatura e pressione alle quali la roccia, da materiale fragile che si frattura sotto sforzo, diventa un mezzo plastico che reagisce alle sollecitazioni deformandosi gradualmente. La pressione e la temperatura aumentano con la profondità, cosicché la roccia diventa plastica alla profondità di circa 60 chilometri (*curva in blu*). La roccia nelle zone dove avvengono terremoti profondi è anormalmente fredda, ma anche se fosse 500 gradi più fredda del mantello «normale», la roccia dovrebbe essere plastica a circa 100 chilometri (*curva in grigio*). Tuttavia si sono registrati terremoti fino alla profondità di quasi 700 chilometri.

Gli eventi profondi registrati negli ultimi 25 anni, differenziati per profondità, sono mostrati in questa carta, che riporta anche i terremoti storici molto intensi (quelli con magnitudo superiore a 7) con ipocentri a profondità particolarmente grandi (più di 630 chilometri), insieme all'anno in cui sono avvenuti. Quasi tutti i terremoti profondi avvengono nei pressi di una fossa oceanica, dove una delle rigide zolle litosferiche che compongono la superficie della Terra si immerge nel mantello nel cosiddetto processo di subduzione.

La distribuzione statistica dei terremoti varia con la profondità. Il numero di terremoti con magnitudo maggiore di 5 in ogni intervallo di 10 chilometri di profondità (*a sinistra*) presenta un minimo a circa 400 chilometri, poi cresce di nuovo per cadere bruscamente a zero a circa 650 chilometri. Le repliche sono rare per la maggior parte degli eventi profondi di intensità moderata, ma divengono più comuni alle massime profondità (*al centro*). I terremoti più forti sono generalmente superficiali, ma a profondità maggiori la magnitudo degli eventi più intensi rimane abbastanza costante, finché l'attività sismica cessa del tutto (*a destra*). I dati suggeriscono che una transizione di fase cristallina delle rocce del mantello, che si pensa avvenga tra 400 e 650 chilometri, possa influenzare i terremoti profondi.

La zona di subduzione è sede di quasi tutti i terremoti profondi. Le profondità degli ipocentri lungo la fossa di Tonga-Kermadec, una fossa oceanica nel Pacifico sudoccidentale ai cui margini si riscontrano attività sismica e isole vulcaniche, corrispondono a una serie di linee approssimativamente parallele di profondità crescente; questo andamento individua una cosiddetta zona di Wadati-Benioff (*in alto*). La zona di Wadati-Benioff segna la zona di subduzione di una zolla litosferica (*in basso*): i terremoti hanno luogo nel lastrone di litosfera in subduzione. Alla flessione verso il basso della zolla è dovuta la formazione della fossa, mentre il materiale fuso che risale dal lembo di litosfera subdotta alimenta la linea di vulcani. L'energia in gioco è fornita dalla circolazione convettiva del mantello.

superiore, cioè al di sopra di questo limite, e il materiale del mantello superiore e di quello inferiore non potrebbero in nessun caso mescolarsi. Il salto di velocità delle onde sismiche alla profondità di 650 chilometri segnerebbe dunque presumibilmente un cambiamento nella composizione del mantello. In alternativa a questa ipotesi (e questo è uno dei punti di maggiore controversia nella geofisica della Terra solida) supponiamo che il lembo di crosta in subduzione possa penetrare nel mantello inferiore. La convezione riguarderebbe allora l'intero mantello e il limite dei 650 chilometri segnerebbe semplicemente una transizione di fase in un mezzo di composizione uniforme. Alla discontinuità nelle proprietà meccaniche in corrispondenza di questo limite sarebbe dovuto il brusco azzeramento dell'attività sismica.

La discesa stessa di una zolla in subduzione fornisce sotto diverse forme l'energia che verrà liberata nei terremoti profondi. Una zolla in discesa viene piegata in vario modo, stirata o compressa; il riscaldamento e le transizioni di fase possono dal canto loro generare tensioni meccaniche, modificando il volume delle rocce. Come vengono liberate le tensioni? Che cosa avviene in realtà nell'ipocentro di un terremoto profondo?

I terremoti profondi possono raggiungere un grado di intensità solo di poco inferiore ai più intensi terremoti superficiali: il terremoto profondo più distruttivo degli ultimi anni, quello rumeno del 1977, aveva magnitudo 7,2 e un terremoto di magnitudo 7,6 avvenne nel 1970 in Colombia, 650 chilometri sotto la superficie. Il meccanismo con cui l'energia viene rilasciata in questi terremoti deve comunque essere diverso dalla fratturazione di rocce a comportamento fragile che dà origine ai terremoti poco profondi. Anche se il materiale nel quale hanno origine i terremoti profondi è molto più freddo - e quindi meno plastico - di quanto avesse pensato Jeffreys, non dovrebbe comunque fratturarsi come le rocce che si trovano a basse pressioni. Se mai le tensioni presenti dovessero aprire una frattura nel lembo di crosta in subduzione, il peso di tutte le rocce sovrastanti la richiuderebbe semplicemente. Quindi, se la roccia è soggetta a deformazione, deve deformarsi plasticamente. L'obiezione di Jeffreys regge ancora: i terremoti «convenzionali», quelli cioè in cui si ha fratturazione e scorrimento di roccia, non possono aver luogo nel mantello.

Gli studi sismografici recenti dei terremoti profondi concordano nel negare la presenza di meccanismi convenzionali. Quasi tutti gli eventi superficiali sono accompagnati da molte scosse di minore entità, le cosiddette repliche. Queste si originano spesso lungo lo stesso piano di scorrimento iniziale, come se rilasciassero una tensione residua lungo la frattura. Le repliche sono molto più rare per gli eventi profondi. Il sisma del 1970 in Colombia, probabilmente il più intenso terremoto molto profondo degli ultimi 25 anni, non ebbe alcuna replica. I terremoti profondi, quando hanno repliche, ne contano una o pochissime.

Le repliche hanno una distribuzione spaziale del tutto differente da quella relativa ai terremoti superficiali. Recentemente Raymond J. Willemann, allora al Los Alamos National Laboratory, e io abbiamo studiato le relazioni spaziali tra le scosse iniziali e le repliche dei terremoti profondi. I piccoli terremoti superficiali presentano spesso repliche con epicentro relativamente vicino a quello dell'evento principale, il che concorda con l'idea secondo la quale le repliche sono l'esito di uno scorrimento lungo la medesima frattura che ha prodotto la scossa principale. Abbiamo comunque trovato che alcuni piccoli terremoti profondi - con magnitudo 5,5 al massimo - presentano repliche a distanze di 30 chilometri o più dalla scossa iniziale. Non è verosimile che zone di frattura che danno origine a terremoti così piccoli si estendano per 30 chilometri.

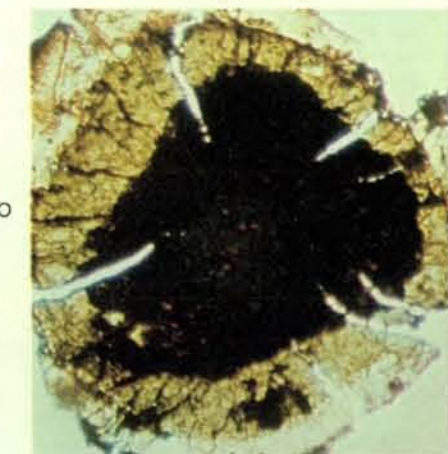
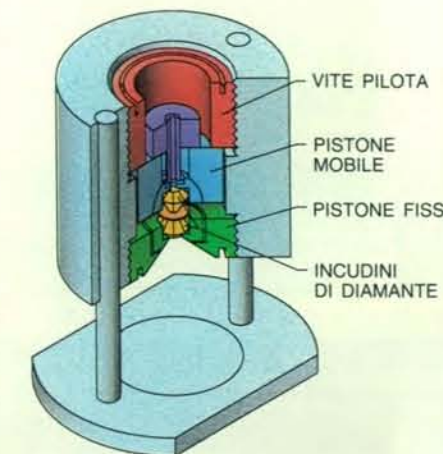
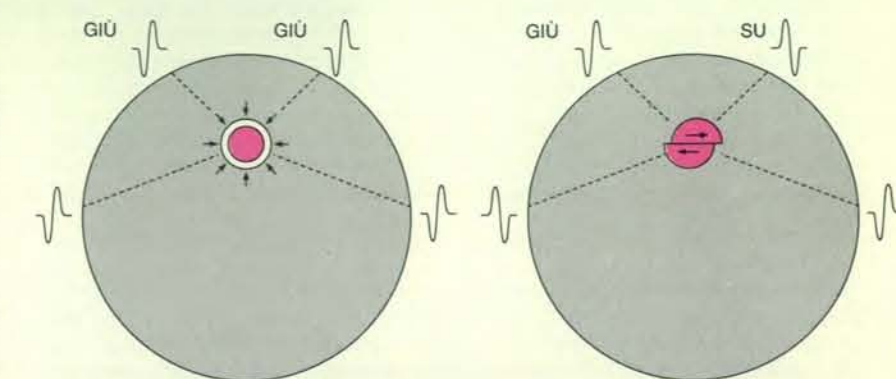
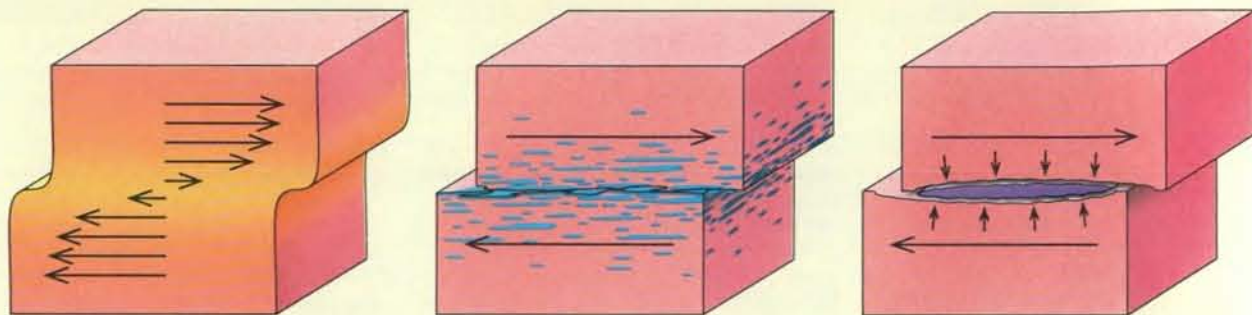
Le repliche dei terremoti profondi, inoltre, non hanno ipocentri disposti lungo un piano come spesso avviene nei terremoti superficiali. I dati disponibili fanno pensare, al contrario, che esse siano distribuite più o meno a caso in uno spazio tridimensionale nell'intorno dell'evento iniziale. Anche questo indica che i terremoti profondi e quelli superficiali hanno meccanismi fondamentalmente differenti.

Un modo accattivante, ma scorretto, di spiegare l'esistenza dei terremoti profondi venne proposto appena dopo la scoperta di questi eventi. Si ipotizzava che essi fossero il diretto risultato della trasformazione di materiale subdotto in una fase più densa. Tali trasformazioni debbono avvenire nella roccia subdotta e, se esse si verificano abbastanza velocemente, la contrazione della roccia dovuta all'aumento di densità potrebbe provocare il rilascio di energia da «implosione» sotto forma di onde sismiche.

Purtroppo, nei rilevamenti, le onde sismiche dei terremoti profondi non assomigliano per nulla a onde da implosione. In un'implosione il materiale si muove verso l'interno, cioè verso il fuoco. Ci si dovrebbe pertanto attendere che tutti i sismografi (che registrano tanto la direzione quanto l'ampiezza delle onde sismiche) segnalino in questi casi un movimento iniziale verso il basso. Inoltre, dato che un'implosione produce un movimento radiale anziché trasversale, dovrebbe generare onde *P* più intense delle onde *S*.

Nella realtà dei fatti il primo movimento originato in superficie da un terremoto profondo è verso il basso in alcune aree e verso l'alto in altre, così come avviene per i terremoti superficiali. I movimenti verso l'alto e verso il basso sono distinti, come se parte della roccia si muovesse in un senso lungo un piano di scorrimento e parte in senso opposto; ciò corrisponde a quanto si verifica per i terremoti superficiali. Inoltre nei terremoti profondi, così come in quelli superficiali, le onde *S* sono molto più intense delle onde *P*, il che porta a identificare come sorgente uno scorrimento piuttosto che un'implosione.

Ma come può la roccia avere scorrimenti improvvisi se le enormi pressioni del mantello impediscono il comportamento «fragile» con fratturazione? Una spiegazione, proposta negli anni sessanta da David T. Griggs dell'Università della California a Los Angeles ed elaborata recentemente da Masaki Ogawa dell'Università di Tokyo, postula una rapidissima deformazione plastica. Un terremoto profondo potrebbe avere luogo quando la roccia che si sta deformando sotto forze di taglio inizia a sviluppare calore di attrito più rapidamente di quanto la roccia circostante riesca a disperderlo. Il calore rende la roccia meno rigida e può perfino fonderla in parte, accelerando così la deformazione. Questo processo di retroazione può far sì che l'aumento della temperatura e della velocità di scorrimento avvengano in modo «esplosivo» e producano un terremoto.

La plausibilità di questo meccanismo dipende in maniera determinante dalla composizione e dalla struttura della roccia in una zona di Wadati-Benioff. Esso viene favorito se la struttura della roccia

Le transizioni di fase nella roccia del mantello, che potrebbero avere un ruolo nei terremoti profondi, possono essere simulate in una cella a incudini di diamante, che comprime campioni di roccia tra due punte di diamante (*a sinistra*). La trasparenza del diamante permette di riscaldare il campione con un fascio laser e di fotografarlo. In un campione di olivina (uno fra i principali costituenti del mantello), compresso a 300 000 atmosfere e riscaldato a 1500 gradi Celsius, le diverse fasi formano anelli concentrici (*a destra*). Un anello esterno chiaro di olivina inalterata passa alla fase di spinello, più gialla, in una transizione che si pensa avvenga alla profondità di 400 chilometri; al centro, dove la pressione e la temperatura sono massime, si vede la fase scura di ossido in cui lo spinello si trasforma a 650 chilometri. La fotografia è stata fornita da William A. Bassett della Cornell University.

L'implosione dovuta alla repentina trasformazione di roccia subdotta in una fase più densa non può essere ritenuta causa dei terremoti profondi. Un'implosione dovrebbe manifestarsi nei sismogrammi con un movimento iniziale verso il basso (*a sinistra*). Invece i terremoti profondi generano un movimento iniziale verso l'alto in alcuni punti e verso il basso in altri (*a destra*), il che indica piuttosto uno scorrimento laterale in rocce a grandi profondità.

I meccanismi di scorrimento proposti concordano nell'ammettere che a grandi profondità la roccia non può semplicemente fratturarsi. In un modello la lenta deformazione della roccia accelera bruscamente quando il calore d'attrito si accumula, rendendo meno rigida la roccia e affrettando la deformazione stessa in modo «galoppante» (*a sinistra*). Un secondo meccanismo attribuisce lo scorrimento all'azione di fluidi (*al centro*). Al di sotto di una certa profondità l'alta pressione può liberare l'acqua contenuta nella struttura cristallina dei minerali; l'acqua può neutralizzare le forze che «contengono» l'apertura di faglie potenziali, permettendo quindi che queste si sviluppino. Secondo una terza proposta gli sforzi di taglio possono indurre una transizione di fase in uno strato parallelo agli sforzi stessi (*a destra*). Il rapido cambiamento della struttura cristallina indebolirebbe la roccia, permettendo lo scorrimento.

- per esempio per la presenza di uno strato debole - tende a far sì che lo scorrimento plastico avvenga lungo un piano. Non è affatto certo che la stratificazione nel materiale subdotto abbia l'orientazione giusta perché lo scorrimento venga favorito lungo le direzioni che si osservano più spesso nei terremoti profondi.

Un modello differente attribuisce lo scorrimento nelle zone in cui avvengono i terremoti profondi all'effetto di fluidi intrappolati. Le prove di laboratorio hanno dimostrato che a pressioni corrispondenti a quelle che si hanno a scarse profondità, i fluidi contenuti nei pori delle rocce possono contrastare le forze che «contengono» una potenziale faglia, facendo sì che questa si apra sotto uno sforzo di taglio inferiore a quello che sarebbe altrimenti necessario. Almeno in un caso (Rocky Mountain Arsenal, presso Denver) si è registrata una sequenza di eventi sismici superficiali successiva all'iniezione nel sottosuolo di residui liquidi che, a quanto pare, abbassano le forze di contenimento, permettendo agli strati rocciosi di scorrere.

Nel 1966 C. B. Raleigh e Mervyn S. Paterson della Australian National University ipotizzarono che la pressione dei fluidi contenuti nei pori della roccia nelle zone dove avvengono i terremoti profondi potesse avere il medesimo effetto, permettendo al materiale subdotto di fratturarsi come la roccia che si trova a pressioni minori. I fluidi potrebbero derivare, secondo Raleigh e Paterson, dalla disidratazione di minerali come il serpentino (una forma di silicato di magnesio) presenti nel materiale subdotto: la liberazione dell'acqua incorporata nella struttura cristallina dei minerali avviene per il riscaldamento a temperature superiori a 500 gradi Celsius a opera del mantello circostante. Raleigh e Paterson proposero anche altre origini per i fluidi: acqua rimasta intrappolata nei sedimenti di fossa oceanica trascinati in profondità assieme alla crosta e roccia del mantello parzialmente fusa.

Per influenzare le proprietà meccaniche di una roccia, come la tendenza alla fratturazione, un fluido dovrebbe potere «migrare» attraverso di essa e non è certo che la roccia del mantello sia abbastanza porosa da permetterlo. Inoltre la disidratazione dei minerali dovrebbe avvenire a determinate temperature e pressioni, tipiche di determinate profondità. Se la pressione dei fluidi contenuti nei pori provoca i terremoti profondi, a queste profondità si dovrebbe riscontrare un'intensa concentrazione di fenomeni sismici. Ma la frequenza dei terremoti presenta solo una moderata variazione con la profondità.

La variazione che si osserva (la caduta di attività sismica ai 400 chilometri del limite olivina-spinello e la sua ripresa a profondità maggiori) sembra avere qualche relazione con le profondità delle transizioni di fase. Stephen H. Kirby dello US Geological Survey ha proposto un meccanismo per i terremoti profondi che dipende dalle transizioni di fase ma, a differenza delle ipotesi precedenti, si attua sotto forma di scorrimenti anziché di implosioni. In sostituzione della reale roccia del mantello Kirby e colleghi hanno studiato il ghiaccio e la tremolite (un silicato di calcio e magnesio), che hanno in comune la proprietà di passare a fasi più dense a pressioni facilmente producibili in laboratorio.

Comprimendo questi materiali a pressioni appena inferiori a quella della normale transizione di fase e sottoponendoli a sforzi di taglio, si è trovato che la transizione di fase ha inizio lungo uno straterello parallelo allo sforzo di taglio. La rapida riorganizzazione della struttura cristallina in questo straterello indebolisce il materiale, permettendo lo scorrimento. Kirby e colleghi hanno notato che durante questo processo i campioni emettevano scricchiolii e crepitii, cioè analoghi di laboratorio dei terremoti.

Kirby ipotizza che simili transizioni di fase premature abbiano pure luogo nella roccia subdotta sottoposta a tensioni e che lo scorrimento risultante spieghi almeno alcuni dei terremoti che avvengono nelle zone di Wadati-Benioff. Questa ipotesi non sembra essere in contrasto con il fatto che i terremoti profondi si verificano in un ampio intervallo di profondità al di sotto del limite di 400 chilometri: molti ricercatori, tra cui William A. Bassett della Cornell University, hanno trovato che le transizioni di fase nel materiale subdotto possono verificarsi a profondità molto diverse, in dipendenza dalla precisa composizione del materiale e dalla velocità con cui esso discende. Il meccanismo di Kirby dovrebbe comunque dar ragione della brusca scomparsa dei terremoti profondi a profondità maggiori di 680 chilometri. A quella profondità tutte le transizioni di fase conosciute per il mantello hanno già avuto luogo.

Non è ancora dimostrato che gli sforzi di taglio abbiano, nelle rocce del mantello, lo stesso effetto sulle transizioni di fase osservato nel ghiaccio e nella tremolite. Ma anche se l'ipotesi di Kirby fosse sbagliata le transizioni di fase possono avere un ruolo nei terremoti profondi. Forse esse generano semplicemente tensioni che sono liberate bruscamente altrove secondo modalità sconosciute.

Sarà forse presto possibile scegliere tra le varie ipotesi con maggiore certezza. Raymond Jeanloz e Charles Meade dell'Università della California a Berkeley riproducono in laboratorio le condizioni presenti nel mantello al fine di studiare i meccanismi proposti per i terremoti profondi. Un piccolo dispositivo, la cosiddetta «cella a incudini di diamante», produce le pressioni necessarie comprimendo un piccolo campione di roccia tra le punte di due diamanti. Il campione può essere riscaldato facendo passare un fascio laser attraverso uno dei due diamanti; le transizioni di fase e le altre variazioni che avvengono nella roccia possono essere osservate attraverso l'altro diamante, mentre appositi sensori acustici rilevano i «terremoti» che si producono. Questi studi sono ancora in fase preliminare, ma già si può desumere che ad alte pressioni l'olivina si indebolisce solo quando contiene anche serpentino, un risultato che depone a favore del meccanismo della disidratazione proposto da Raleigh e Paterson.

Dimostrando rispettivamente la realtà dei terremoti profondi e la loro «impossibilità», Wadati e Jeffreys (che sono entrambi viventi a distanza di più di 60 anni) hanno posto un problema sulla cui soluzione i geofisici si stanno ancora arrovellando. Nel contesto della tettonica a zolle e della convezione sublitosferica, che i terremoti profondi stessi hanno contribuito a far comprendere, le caratteristiche di questi eventi sismici hanno aperto nuovi enigmi.

Uno riguarda il limite di 680 chilometri, sotto il quale non avvengono più terremoti: si trova in corrispondenza del limite inferiore della convezione nel mantello o semplicemente segna una variazione nelle proprietà meccaniche di un mantello che presenta convezione anche a profondità maggiori? I terremoti profondi che avvengono in regioni in cui è certo che non vi sono zone di subduzione costituiscono un ulteriore enigma. I terremoti profondi in Romania e nell'Hindu Kush, due di queste regioni, possono essere la manifestazione di vecchie zone di subduzione messe in ombra da un'attività tettonica più recente. Questa spiegazione è meno plausibile per i movimenti tellurici registrati talvolta in Africa settentrionale e in Spagna. Qui l'enigma dei terremoti profondi è avvolto da un ulteriore mistero: la possibilità che alcuni di essi possano avvenire in completa assenza di subduzione.

BIBLIOGRAFIA

WADATI K., *Shallow and Deep Earthquakes* in «The Geophysical Magazine», 1, n. 4, marzo 1928.

LEITH ANDREW e SHARPE J. A., *Deep-Focus Earthquakes and Their Geological Significance* in «The Journal of Geology», 44, n. 8, novembre-dicembre 1936.

FROHLICH CLIFF, *Kiyoo Wadati and Early Research on Deep Focus Earthquakes: Introduction to Special Section on Deep and Intermediate Focus Earthquakes* in «Journal of Geophysical Research», 92, n. B13, 10 dicembre 1987.

KIRBY STEPHEN H., *Localized Polymorphic Phase Transformations in High-Pressure Faults and Applications to the Physical Mechanism of Deep Earthquakes* in «Journal of Geophysical Research», 92, n. B13, 10 dicembre 1987.

La larva di *Drosophila* si colora di blu in corrispondenza della parte anteriore dell'intestino medio (*in alto*) e dei corpi grassi (*in basso*), indicando così che nel metabolismo degli zuccheri è interessata l'attività di un enzima. Il gene per l'enzima è stato attivato da una proteina regolatrice che si trova nelle cellule di lievito, ma che funziona anche nelle cellule di insetti, mammiferi e piante.

Come funzionano gli attivatori dei geni

La modalità con cui i geni vengono attivati e disattivati è ben nota per i batteri, ma studi di biologia molecolare hanno dimostrato che il medesimo meccanismo agisce nella regolazione genica degli organismi superiori

di Mark Ptashne

Ogni cellula umana contiene circa 100 000 geni, ma in ogni dato momento solo una frazione di questi è attiva. Molti geni si esprimono in modo selettivo, per esempio solo durante certi stadi dello sviluppo o in risposta a segnali provenienti dall'ambiente. Come avviene la loro attivazione o disattivazione? Sono ormai 20 anni che i miei colleghi e io, alla Harvard University e in altre sedi, ci dedichiamo allo studio del meccanismo di attivazione in un virus batterico, il batteriofago lambda. Ci accorgiamo ora che, in effetti, i concetti che abbiamo sviluppato nel corso di quegli studi ci hanno portato lontano; ora possiamo applicarci all'analisi della regolazione genica nelle cellule umane e in altri organismi superiori.

Per capire il problema, si ricordi che un gene tipico, che è poi un segmento di DNA, codifica per una proteina. Questo significa che la sequenza specifica delle coppie di basi che costituiscono quel segmento corrisponde alla specifica sequenza di amminoacidi che costituisce la proteina. La decifrazione di un gene richiede, innanzitutto, che esso sia trascritto in una molecola di RNA messaggero (m-RNA) e quindi che questa molecola sia tradotta in proteina.

L'espressione genica può essere controllata da proteine regolatrici che si legano a siti specifici del segmento di DNA. Queste proteine sono dette attivatori o repressori a seconda che accelerino o rallentino l'attività di trascrizione; alcune possono eseguire ambo le funzioni. In che modo queste proteine riconoscono i siti specifici sul DNA e attivano o disattivano i geni da esse controllati?

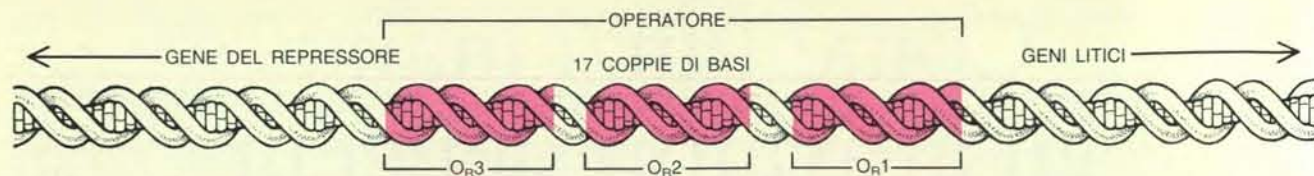
Una parte della risposta chiama in causa particolari enzimi, le RNA-polimerasi, responsabili della trascrizione dei geni in RNA. Nel caso dei geni di cui ci occuperemo in questo articolo, una RNA-polimerasi si attacca al DNA vicino all'inizio di un gene e quindi si sposta lungo quel gene trascrivendo la sequenza di DNA in una molecola di m-RNA. Ora la domanda può essere posta in modo più preciso: in che modo un attivatore collabora con l'RNA-polimerasi nella trascrizione di un gene? Come fa il repressore a ostacolare la trascrizione da parte dell'enzima?

Nei nostri studi sul fago lambda abbiamo ottenuto numerose informazioni sulle interazioni tra regolatori, DNA e RNA-polimerasi, che spiegano come avvenga la regolazione genica in quel semplicissimo organismo. Ma avevamo molti motivi per sospettare che i risultati ottenuti non avessero attinenza con quanto avveniva negli organismi superiori. Tanto per cominciare, le RNA-polimerasi negli organismi superiori e inferiori esistono in forme che, a prima vista, sembrano diverse. E mentre i regolatori che abbiamo studiato nel fago lambda si legano molto vicino al gene che controllano, in altri organismi si legano sul DNA in siti che si trovano a distanza di centinaia, o anche migliaia, di coppie di basi dal gene. Inoltre, negli eucarioti (organismi superiori), ma non nei procarioti come i batteri, il DNA è segregato all'interno di un nucleo cellulare ed è avvolto attorno a proteine particolari, gli istoni, come un filo è avvolto su un rocchetto. In questo caso intervengono meccanismi fondamentalmente diversi?

I progressi compiuti negli ultimi due anni suggeriscono che alcuni semplici principi riguardanti la regolazione genica possono essere comuni anche alle situazioni più disparate (batteri, lieviti, piante, moscerini della frutta o esseri umani). Esporrò questi concetti così come sono stati sviluppati nel corso degli studi sul fago lambda e successivamente descriverò come i miei colleghi e io li applichiamo allo studio della regolazione genica negli organismi superiori. Considererò in particolare il problema dell'attivazione genica.

Il fago lambda ha attratto l'attenzione degli scienziati più di trent'anni fa, quando essi si sono resi conto che il ciclo vitale di quel virus è un clamoroso esempio dell'importanza della regolazione genica. Quando in un batterio viene inoculato DNA virale, questo segue l'una o l'altra di due vie: o la maggior parte dei geni virali viene espressa e il virus si replica facendo scoppiare (lisare) la cellula ospite, oppure quasi tutti i geni virali sono disattivati e il DNA virale inerte diventa una parte passiva del cromosoma batterico. Tuttavia, quando il batterio è in presenza di un agente cancerogeno, per esempio di radiazioni ultraviolette, i geni silenti vengono attivati.

Una proteina regolatrice, codificata dal virus, svolge un ruolo chiave nel controllare l'espressione dei geni virali. Denominata repressore lambda, essa si lega ad alcuni siti specifici del cromosoma virale e disattiva la maggior parte dei geni virali. Gli agenti cancerogeni possono indurre l'espressione dei geni per la replicazione virale e la lisi cellulare (geni «litici») distruggendo il repressore. Ma, oltre a essere un repressore dell'espressione genica, come indica il suo nome, la proteina *attiva* anche la trascrizione del proprio gene, ovvero la fa avvenire più velocemente.

La conoscenza della struttura del repressore ci ha aiutato a capire in che modo esso funzioni. L'unità fondamentale, o monomero, del repressore è ripiegata in due «globi» (o domini) di dimensione pressoché uguale. Due monomeri si associano in un dimero, che è la forma del repressore che si lega al DNA. Sul DNA del fago lambda, due dimeri si legano a siti adiacenti dell'operatore, l'O_R1 e l'O_R2 (O sta per operatore e R per *right*, destra); vicino vi è un altro sito, l'O_R3,

L'operatore lambda contiene i siti in corrispondenza dei quali si legano le proteine regolatrici che controllano i geni di un virus batterico, il fago lambda. In questo schema, il gene per una proteina regolatrice, detta repressore, si trova a sinistra dell'operatore; a destra di questo vi sono i geni «litici», che mediano la replicazione del virus e la lisi della cellula batterica ospite. All'interno dell'operatore vi sono tre siti di legame, O_R1, O_R2 e O_R3, ciascuno dei quali è costituito da un segmento di DNA di 17 coppie di basi.

di cui non parleremo perché non interviene nella repressione (*si vedano le illustrazioni in questa pagina*).

La presenza di questi due dimeri legati al DNA influisce sulla trascrizione in due modi, l'uno negativo e l'altro positivo. In primo luogo, i repressori impediscono alla RNA-polimerasi di legarsi al DNA e di copiare i geni litici (che, per convenzione, si trovano a destra). In secondo luogo, uno dei repressori aiuta la RNA-polimerasi a unirsi al DNA e a cominciare la trascrizione del gene che codifica per il repressore stesso (si è stabilito che questo gene è situato a sinistra dei repressori legati).

I siti ai quali si lega il repressore sono chiamati operatori e quelli a cui si lega la RNA-polimerasi promotori. Il promotore a destra dell'operatore lambda ha un'azione energica, mentre quello a sinistra ha un'azione debole. Gli operatori ai quali si lega il repressore si sovrappongono al promotore «energico», che si trova sulla destra, mentre sono contigui a quello «debole», che si trova sulla sinistra. In assenza di repressore, la RNA-polimerasi si attacca al promotore energico e dà inizio alla trascrizione. Per contro, essa si attacca al promotore debole e trascrive il gene del repressore solo quando è coadiuvata dal repressore adiacente legato (che svolge in questo caso il ruolo di attivatore). Una lieve differenza nelle sequenze dei due promotori spiega perché un promotore attragga solo debolmente la polimerasi e richieda così un attivatore per poter funzionare, mentre l'altro funzioni bene fino a quando la polimerasi è in grado di accedere a esso.

Vi sono valide ragioni biologiche alla base della complessità di un simile quadro, con i due dimeri del repressore che si legano simultaneamente e il duplice effetto positivo-negativo del repressore. Questi e altri fattori permettono al virus di passare da una modalità di sviluppo all'altra in funzione dell'ambiente (si veda l'articolo *Un interruttore genetico in un virus batterico* di Mark Ptashne, Alexander D. Johnson e Carl O. Pabo in «Le Scienze» n. 173, gennaio 1983). Ma, a questo punto, ritorno alla domanda originale così modificata: in che modo esattamente il repressore si lega ai siti operatori specifici e attiva la trascrizione del proprio gene?

La cristallografia a raggi X e un gran numero di esperimenti biochimici hanno permesso di avere una visione abbastanza particolareggiata di come il repressore lambda e le proteine regolatrici a esso correlate si leghino selettivamente ai loro siti operatori sul DNA. La superficie che si lega al DNA di ogni monomero del repressore presenta una struttura sporgente, chiamata elica alfa. Eliche alfa si trovano in molte proteine e numerose sequenze diverse di amminoacidi si ripiegano in questo modo.

L'elica alfa che sporge da ogni monomero si inserisce bene nel solco principale che percorre l'intera doppia elica di DNA. I gruppi chimici situati alla superficie esterna dell'elica danno luogo a una configurazione determinata dalle sequenze amminoacidiche che si susseguono lungo di essa. Anche i gruppi chimici che si trovano sulle coppie di basi esposte nel solco principale danno luogo a una configurazione determinata dalla sequenza di tali coppie. Il repressore si può legare solo quando le due configurazioni combaciano (*si veda l'illustrazione a pagina 29 in alto*).

Il fatto che il repressore si leghi come dimero significa che ciascuna delle due eliche alfa identiche (che sono chiamate in questo contesto eliche di riconoscimento) deve trovare il proprio corrispondente nella sequenza di DNA. Questa esigenza spiega perché ogni sito operatore contenga due mezzi siti identici, o quasi identici, ciascuno dei quali viene riconosciuto da una delle due eliche alfa. Il dimero repressore ha una simmetria doppia e la stessa cosa vale per il suo operatore.

Il repressore lambda è in grado di individuare, in mezzo a una grande quantità di DNA, una sequenza specifica di coppie di basi e, quando si lega, non deforma in maniera significativa la doppia elica del DNA. Molte altre proteine, sia nei procarioti sia negli eucarioti, riconoscono il DNA mediante un meccanismo analogo; le loro eliche di riconoscimento differiscono per la sequenza da quella del repressore lambda, così come diverse sono le sequenze degli operatori ai quali si legano. Anch'esse riescono a identificare un sito di legame avente una lunghezza tipicamente inferiore a 20 coppie di basi in mezzo a milioni di sequenze che non hanno alcuna relazione con esso. (Per inciso, ciascuna di queste proteine, repressore incluso, presenta una seconda elica alfa sporgente, adiacente all'elica di riconoscimento, e la coppia forma una caratteristica struttura a «bielica». La seconda elica alfa è disposta trasversalmente al solco principale del DNA e contribuisce a orientare l'elica di riconoscimento all'interno di tale solco.)

Questa rappresentazione dell'interazione del repressore con il suo bersaglio lungo la molecola del DNA ha influenzato notevolmente le nostre ipotesi sul funzionamento degli attivatori dei geni. Per esempio, si potrebbe immaginare che il repressore, legandosi al DNA, costringa i due filamenti a separarsi, facendoli poi riavvolgere in una struttura insolita. In questo quadro è concepibile che una struttura di DNA alterata a tal punto possa in qualche modo essere responsabile dell'attivazione di un gene. Ma il fatto che il DNA non modifichi notevolmente la propria forma quando il repressore lambda si lega a esso suggerisce che l'attivazione della trascrizione non sia da attribuire a tale legame di per sé.

In che modo, allora, il repressore legato attiva la trascrizione? Una sua parte, distinta dalla superficie con la quale esso si lega al DNA, interagisce con la RNA-polimerasi per attivare la trascrizione. Questa «superficie attivante» entra in contatto con l'enzima e ne favorisce il legame, avviando così la trascrizione del gene del repressore. Sappiamo ancora poco sulla struttura della RNA-polimerasi, mentre abbiamo qualche informazione in più sulla superficie attivante del repressore.

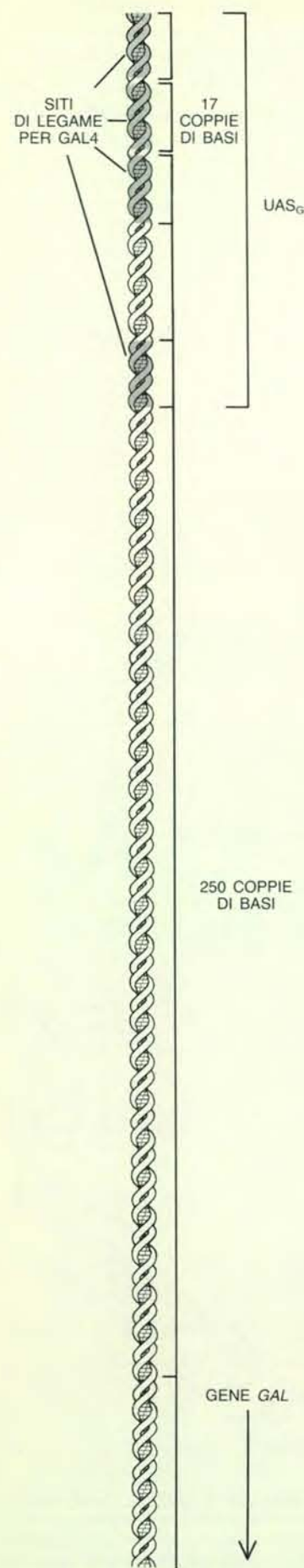
Riteniamo che un elemento importante della superficie attivante del repressore lambda sia anch'esso un'elica alfa. Abbiamo tratto questa conclusione dall'analisi di una classe speciale di mutanti del repressore, che si legano normalmente al DNA, ma non ne attivano la trascrizione. Mutanti di questo tipo provocano cambiamenti in un'elica alfa, che potremmo chiamare elica di attivazione, o vicino a essa. Esaminando un modello di repressore legato in prossimità della RNA-polimerasi si vede che l'elica è posizionata esattamente in maniera tale da poter entrare in contatto con la polimerasi.

Le sostituzioni di amminoacidi nei repressori mutanti che si legano al DNA, ma non ne attivano la trascrizione appartengono a un tipo particolare: diminuiscono tutte la quantità di carica negativa presente nell'elica di attivazione o vicino a essa. La maggior parte degli amminoacidi che costituiscono le proteine è elettricamente neutra: i pochi amminoacidi dotati di carica positiva o negativa sono presenti nella maggior parte delle proteine in numero grosso modo uguale. I dati ricavati dai nostri mutanti lambda suggeriscono che gli amminoacidi con carica negativa siano particolarmente importanti per l'attivazione, un'ipotesi che ricorrerà spesso negli esperimenti di cui parlerò più avanti.

C'è un altro aspetto della modalità d'azione del repressore che risulta essere di grande importanza: esso ha a che fare con il legame dei due dimeri del repressore, indispensabile per l'attivazione. I dimeri non si legano indipendentemente, ma piuttosto il legame dell'uno favorisce quello dell'altro. Per capire come ciò possa avvenire, si immagini che un dimero del repressore si leghi in primo luogo con il sito più a destra (O_R1), quello che è noto per essere il sito di legame più forte. Questo repressore legato aiuta quindi un altro dimero a legarsi al sito adiacente, entrando in contatto con esso. Questo tipo di legame viene detto «cooperativo».

È chiaro ora che il principio del legame cooperativo è ampiamente utilizzato dalle proteine che si legano al DNA. L'approfondimento del fenomeno va oltre i limiti di questo articolo; in poche parole, riteniamo che la cooperazione renda il legame delle proteine estrema-

Il repressore attiva il proprio gene aiutando un enzima, la RNA-polimerasi, a legarsi a esso. La RNA-polimerasi è l'enzima che trascrive i geni in RNA, prima fase del processo mediante il quale un gene si esprime come proteina. I «dimeri» del repressore si legano sia a O_R1 sia a O_R2. (Un dimero è una molecola che consta di due subunità identiche.) Si ritiene che una parte della proteina, la «superficie di attivazione», entri in contatto con l'enzima. I dimeri della proteina si «toccano» anche l'uno con l'altro per facilitare il legame reciproco, un fenomeno noto come cooperazione. I dimeri legati bloccano, inoltre, la trascrizione dei geni litici situati alla loro destra, impedendo il legame dell'RNA-polimerasi; è per questo che alla proteina è stato attribuito il nome di «repressore».

La cooperazione a distanza avviene quando i siti di legame per il repressore sono separati. Il DNA intercalare forma un'ansa che consente ai dimeri di entrare in contatto e quindi di collaborare nella formazione del legame. Secondo l'autore, l'ansa potrebbe anche agevolare le interazioni tra le proteine coinvolte nella regolazione genica degli organismi superiori.

mente sensibile a piccoli cambiamenti nella loro concentrazione, permettendo così ai geni di attivarsi e disattivarsi con la massima efficienza. La cooperazione aiuta anche una proteina a distinguere il sito di legame giusto sul DNA in mezzo a una grande quantità di DNA che non ha alcuna rilevanza per il legame. Nel caso del repressore lambda, l'effetto cooperativo è relativamente modesto: la presenza di un dimero legato fa aumentare la possibilità di legame di un altro dimero di un fattore 10 circa. Ma anche questo piccolo effetto ha un'importanza critica per l'azione del repressore.

Un altro aspetto della cooperazione nel repressore lambda ha influenzato le nostre ipotesi su come le proteine regolatrici possano influire sulla trascrizione anche quando sono legate a notevole distanza dai geni che controllano. Proprio come un repressore può aiutare un altro repressore a legarsi quando i siti che fungono da operatori sono adiacenti, così i dimeri del repressore possono aiutarsi reciprocamente anche se i siti sono distanti lungo la molecola di DNA. I dimeri possono entrare in contatto l'uno con l'altro perché il DNA è flessibile e, per favorire l'interazione, forma un'ansa (*si veda l'illustrazione nella pagina precedente*). L'ipotesi di un DNA in grado di formare anse è stata avanzata da Sankar L. Adhya e collaboratori del National Cancer Institute e da Robert F. Schleif e collaboratori della Brandeis University.

La cooperazione a distanza è stata dimostrata o desunta in molti casi, sulla base di valide prove, per le proteine regolatrici dei procarioti. In effetti il fago lambda costituisce l'eccezione, nel senso che i suoi operatori, nella configurazione ordinaria, sono adiacenti sul DNA. Sembra ragionevole ammettere che lo stesso meccanismo di formazione dell'ansa, ipotizzato per avvicinare le proteine regolatrici, possa anche permettere alle proteine regolatrici legate in un determinato sito di influire sulla trascrizione di un gene distante. Per esempio, un attivatore legato al DNA potrebbe funzionare entrando in contatto con la RNA-polimerasi stessa o qualche proteina accessoria fissata nel punto in cui ha inizio il gene, mentre il DNA intercalare si piega a formare un'ansa per rendere possibile il contatto.

Il gene *GAL* del lievito è controllato da una sequenza attivante a monte (UAS_G), localizzata a una distanza di circa 250 coppie di basi. Esso è così chiamato perché codifica per un enzima che degrada il galattosio; la proteina regolatrice che ne attiva la trascrizione è denominata GAL4. Ciascuno dei quattro siti di legame per GAL4 contenuti in UAS_G ha, come il sito di legame per il repressore, una lunghezza corrispondente a 17 coppie di basi. GAL4 funziona anche quando i suoi siti di legame si trovano a una distanza di 750 coppie di basi dal gene *GAL*.

Come possono i principi che descrivono l'azione del repressore lambda applicarsi ai problemi della regolazione genica negli organismi superiori? È stato piuttosto difficile rispondere a questa domanda perché gli eucarioti sono più difficili da manipolare dei batteri. Anche se lo sviluppo delle tecnologie comunemente note come «tecnologie del DNA ricombinante» ha permesso di realizzare nuovi tipi di esperimenti con le cellule degli organismi superiori, le limitazioni rimangono scoraggianti in confronto a quanto è possibile fare con le cellule batteriche. Abbiamo scelto di studiare i geni nei lieviti, un gruppo di eucarioti tra i più semplici esistenti in natura. I lieviti crescono quasi altrettanto velocemente dei batteri e si possono manipolare dal punto di vista genetico con una facilità paragonabile.

I nostri esperimenti si concentrano su un gruppo di geni che codificano per enzimi che degradano il galattosio. I geni *GAL*, come vengono chiamati, sono attivati dalla proteina GAL4, la quale si lega a una sequenza del DNA che dista all'incirca 250 coppie di basi dall'inizio del più vicino gene *GAL*. Tale sequenza viene indicata con la sigla UAS_G, dall'inglese *galactose upstream activating sequence*, sequenza che attiva a monte il galattosio. L'espressione «a monte» sottolinea il fatto che la sequenza media a distanza l'attività di GAL4. In effetti, la UAS_G può essere posizionata di fronte ad altri geni del lievito, e a varie distanze, e GAL4 attiverà anche la trascrizione di questi geni.

Osservando con maggiore attenzione l'UAS_G, si nota che essa contiene quattro sequenze simili, che si legano a GAL4; ciascuna consta di 17 coppie di basi e, come l'operatore lambda, ha una simmetria doppia (o quasi). Questa constatazione è conforme ai risultati di altri esperimenti, che suggeriscono la possibilità che GAL4, come il repressore lambda, si leghi a ogni sito in forma di dimero. Un dimero di GAL4 legato a un unico sito attiverà la trascrizione, ma in modo non altrettanto efficiente dei dimeri multipli legati a siti multipli.

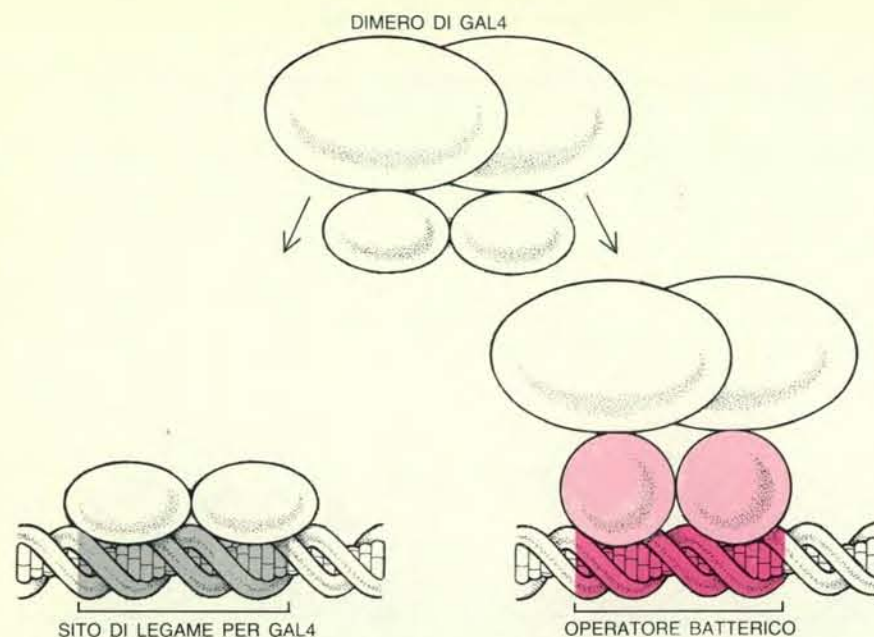
Come funziona GAL4? Vale a dire, come si lega selettivamente a UAS_G per poi attivare la trascrizione di un gene che si trova a una distanza di diverse centinaia di coppie di basi? I seguenti esperimenti dimostrano che, come il repressore lambda, GAL4 ha una porzione della superficie che si lega al DNA e una che ha funzione attivante. Si osserva anche che, diversamente dalle due superfici del repressore lambda (che sono vicine l'una all'altra in un unico dominio), le superfici di legame con il DNA e di attivazione di GAL4 si trovano in differenti parti di questa proteina molto più voluminosa e si separano facilmente.

L'esperimento che ha localizzato la superficie che lega il DNA in GAL4 è stato realizzato in parte frammentando il gene che codifica per questa proteina, reinserendone i frammenti nel lievito e quindi determinando quali funzioni, se ve ne sono, sono svolte dai segmenti proteici prodotti da tali frammenti. Abbiamo trovato un frammento, la parte iniziale della proteina, che si lega al DNA ma non riesce ad attivare l'espressione genica. Questa osservazione ci ha fatto supporre che la superficie di attivazione di GAL4 debba trovarsi in una parte della proteina che non comprende i primi 100 amminoacidi.

L'esperimento che ha dato conferma a questa supposizione ha comportato la fusione di frammenti di DNA provenienti dal gene di GAL4 con frammenti di DNA che codificano per parti di altre proteine, e quindi la produzione di proteine ibride. Nell'esperimento risolutivo abbiamo sostituito la regione che si lega con il DNA in GAL4 con l'analoga regione di un repressore batterico. Come era prevedibile, la proteina ibrida non ha alcun effetto sull'espressione genica nel lievito normale, perché non ha alcun mezzo per riconoscere l'UAS_G. Se, però, un operatore noto per essere riconosciuto dal repressore batterico viene posto in presenza di un gene di lievito, la proteina ibrida è in grado di attivare la trascrizione di quel gene. Il repressore batterico nativo non può attivare l'espressione genica nel lievito. In altre parole, perché un attivatore possa essere tale sono necessarie due funzioni: una superficie di attivazione (in questo caso, il frammento di GAL4) e una superficie di legame con il DNA (in questo caso il repressore batterico), per fissare il frammento attivante al DNA vicino al gene.

Per inciso, non sappiamo con precisione come GAL4 riconosca i propri siti sul DNA; essa non possiede, infatti, alcuna struttura a bielica come il fago lambda. La sua sequenza suggerisce che possa avere sporgenze digitiformi, costituite da amminoacidi e ancorate al resto della struttura da atomi di zinco. L'esperimento che ho appena descritto, però, mostra che questo aspetto, di per sé interessante, non è fondamentale per farci comprendere come avviene l'attivazione dei geni. L'esperimento di ibridazione lievito-batterio ci ha spinto a chiederci in che cosa consistesse la superficie di attivazione di GAL4. Il modo più ovvio per affrontare il problema è stato suggerito dal precedente esperimento con il repressore lambda: l'isolamento di mutanti di GAL4 che si legano al DNA, ma non sono in grado di attivarlo. Ammettendo che questi mutanti siano analoghi ai mutanti del repressore, essi mostrerebbero cambiamenti nella composizione degli amminoacidi, che definirebbero la superficie di attivazione. Ma, per ra-

La superficie del repressore che si lega al DNA interagisce con questo mediante eliche alfa sporgenti, formate da catene di amminoacidi avvolte a elica. Le eliche sono rappresentate qui da corti cilindri. L'«elica di riconoscimento» si inserisce perfettamente nel solco principale che decorre lungo l'elica del DNA come se fosse il filetto di una vite. L'altra elica alfa è trasversale rispetto al solco. La disposizione degli amminoacidi nelle eliche di riconoscimento coincide con quella delle coppie di basi nell'operatore a cui il repressore si lega.

I derivati di GAL4 hanno permesso di identificare le parti della proteina interessate nell'attivazione. Nella proteina intatta (*a*), due regioni attivanti (*in colore*) inducono la trascrizione (*freccia*); senza di esse la trascrizione non avviene (*b*). Ciascuna regione può attivare la trascrizione in modo indipendente quando si attacca alla regione di legame con il DNA (*c*, *d*). Insieme, esse hanno un'efficienza quasi pari a quella della proteina intatta, anche se l'80 per cento di questa manca (*e*). Una catena di 15 amminoacidi, che probabilmente forma un'elica alfa, funziona bene quasi quanto una delle superfici attivanti quando si unisce alla regione di legame con il DNA (*f*).

Manipolando GAL4 si nota che le superfici di legame con il DNA e di attivazione si trovano in parti diverse della proteina. La GAL4 originaria ha almeno due domini (*in alto*); una GAL4 mutante priva del dominio maggiore si lega al DNA, ma non ne attiva la trascrizione (*a sinistra*). Che il dominio contenga la superficie attivante è indicato da esperimenti in cui la regione di legame con il DNA di GAL4 è sostituita dalla regione di legame con il DNA di una proteina batterica correlata con il repressore lambda (*a destra*). L'ibrido attiva un gene quando un sito di legame per il repressore batterico è inserito nel DNA del lievito.

gioni sulle quali ritornerò più avanti, i nostri tentativi di isolarli hanno prodotto soltanto frammenti proteici in cui mancava la maggior parte della proteina. Questi mutanti non ci hanno così aiutato a capire meglio la natura della superficie stessa.

La successiva strada che abbiamo tentato ci è stata suggerita da Keith R. Yamamoto e collaboratori dell'Università della California a San Francisco, i quali stavano frammentando un gene che codifica per una proteina regolatrice umana, recettore di glucocorticoidi. Abbiamo così deciso di frammentare ulteriormente il gene per GAL4 e quindi di attaccare la parte che codifica per la superficie di legame con il DNA a frammenti della rimanente parte del gene. Volevamo ottenere in questo modo una forma semplificata di GAL4, nella quale la superficie di attivazione era attaccata direttamente al frammento che lega il DNA. In realtà, abbiamo trovato due parti della proteina, ciascuna costituita da circa 100 amminoacidi: entrambe possono attivare indipendentemente l'espressione genica quando sono attaccate a un frammento che si lega al DNA. Quando sono presenti tutti e due i frammenti, la proteina è in grado di attivare l'espressione genica quasi come GAL4, anche se circa l'80 per cento di essa è stato eliminato.

In questi risultati vi sono due aspetti sorprendenti. Il primo è che l'esperimento sia stato possibile: coloro che si occupano di biologia molecolare sono soliti pensare che le proteine siano strutture tridimensionali definite in modo piuttosto preciso, e che non possano essere rappezzate facilmente. Una cosa è attaccare parti di una proteina a un'altra e pensare che l'ibrido funzioni (così perlomeno credevamo); un'altra è eliminare ampie parti di una proteina e pensare che essa mantenga la propria attività.

Forse non dovremmo essere tanto sorpresi. Negli ultimi anni, è apparso sempre più evidente che nelle proteine i domini possono essere costituiti da unità funzionali indipendenti. La struttura di GAL4 non è stata ancora determinata, ma il fatto che le regioni attivanti e la funzione di legame del DNA possano essere così facilmente separate e ricombinate sta a indicare che esse si trovano in domini separati.

Il secondo aspetto sorprendente della nostra analisi delle regioni attivanti di GAL4 emerge dal confronto delle loro sequenze. Molto spesso parti di proteine che svolgono analoghe funzioni hanno sequenze e strutture tridimensionali simili. Tuttavia, in questo caso, abbiamo notato solo un elemento comune: la presenza di un grande eccesso di amminoacidi con carica elettrica negativa. Kevin Struhl e collaboratori della Harvard Medical School avevano dimostrato che la regione attivante di un altro attivatore della trascrizione del lievito, il GCN4, ha un eccesso di cariche negative; a parte ciò, la sua sequenza non ha alcuna evidente relazione con l'una o l'altra regione attivante di GAL4.

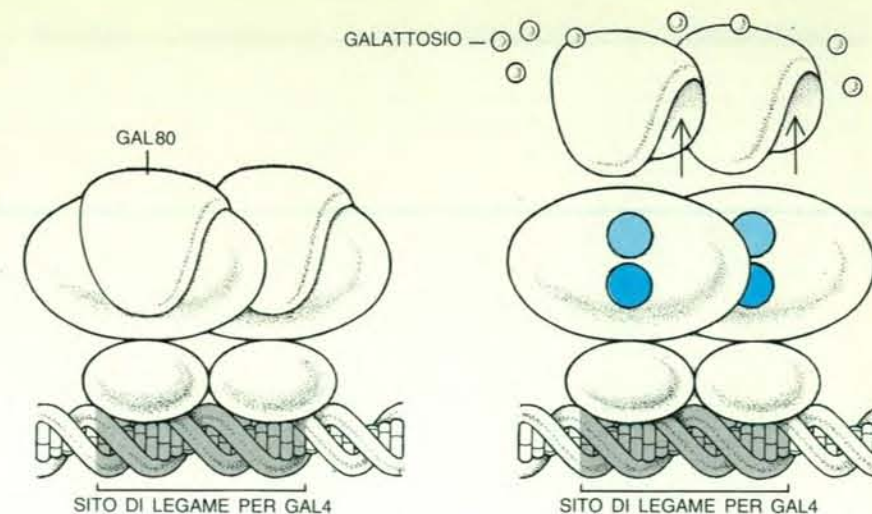
È molto facile trovare sequenze proteiche che abbiano una funzione attivante quando sono unite a un frammento di proteina che si lega con il DNA. Abbiamo attaccato frammenti casuali di DNA batterico a quella parte del gene di GAL4 che codifica per la regione che si lega con il DNA. Abbiamo in seguito introdotto questi geni ibridi nel lievito e abbiamo esaminato quest'ultimo per vedere se le proteine ibride sintetizzate riuscivano ad attivare un gene di prova che si trovava accanto a un sito di legame per GAL4.

Catene di amminoacidi sono state attaccate sperimentalmente al sito di legame con il DNA di GAL4 per analizzare le caratteristiche indispensabili per una superficie di attivazione. Gli stessi amminoacidi compaiono in entrambe le catene, ma in un ordine diverso. Essi sono stati disposti in modo che, se le catene formano eliche alfa (come è mostrato qui), una catena avrebbe gli amminoacidi con carica negativa allineati su una superficie (*in alto*), mentre le cariche negative sull'altra catena sarebbero sparse (*in basso*). La catena con le cariche allineate attiva la trascrizione, mentre la catena con le cariche sparse no. Pertanto, la distribuzione di carica, e non solo la carica stessa, influisce sulla capacità di attivazione.

Un'elevata percentuale delle proteine neoformate attivava l'espressione genica. Inoltre, le loro sequenze non avevano somiglianze cospicue tranne per il fatto che tutte mostravano un eccesso di amminoacidi dotati di carica elettrica negativa. Alcuni dei nuovi attivatori, le cui regioni attivanti hanno una lunghezza corrispondente a soli 50 amminoacidi, funzionano in maniera quasi altrettanto efficiente della proteina GAL4 intatta.

La presenza di cariche negative in eccesso su tanti attivatori suggerisce che ciò non sia frutto di una pura coincidenza; un altro esperimento genetico conferma quest'idea. Ho ricordato prima che i tentativi per definire la regione attivante di GAL4, fatti isolando mutanti che mancavano in modo specifico della funzione di attivazione, hanno fornito solo frammenti della proteina. Se in GAL4 esistono due regioni attivanti ed esse possono funzionare in maniera indipendente, allora per perdere la funzione attivante un mutante dovrebbe modificarsi in almeno due parti della proteina, un evento che potrebbe essere estremamente raro. Solo se la maggior parte della proteina andasse perduta entrambe le regioni attivanti sarebbero assenti.

Partendo, invece, da una versione semplificata di GAL4, possono venire isolate proteine mutanti che hanno un'azione attivante più o meno efficace della proteina originaria. Si nota che vi è una correlazione buona (ma non perfetta) tra carica elettrica e attività: i mutanti con un'attività accresciuta generalmente presentano sostituzioni di amminoacidi che fanno aumentare la carica negativa, mentre i mutanti con un'attività ridotta generalmente presentano sostituzioni di amminoacidi che fanno diminuire la carica negativa.

Tuttavia alcune eccezioni suggeriscono che, oltre alla quantità di carica negativa che la superficie attivante espone, sia importante qualche aspetto strutturale. Ispirati da quanto avevamo scoperto sulla superficie di attivazione del repressore lambda, abbiamo deciso di verificare se un'elica alfa con cariche negative su una delle sue superfici potesse avere un'azione attivante. Abbiamo preparato un frammento di DNA che codifica per un corto segmento di proteina che, in teoria, potrebbe avvolgersi in un'elica di questo tipo e l'abbiamo attaccato al DNA che codifica per quella parte di GAL4 che si lega al DNA. Se l'elica alfa si forma realmente, essa dovrebbe possedere non solo una superficie dotata di carica negativa, ma anche una superficie idrofoba (lipidica), che potrebbe avere lo scopo di favorirne l'adesione alla superficie del frammento di GAL4.

Il galattosio controlla la proteina GAL4 mediante un'altra proteina, la GAL80. Quando esso non si trova nell'ambiente in cui è presente una cellula di lievito, GAL80 scherma le superfici di attivazione di GAL4, impedendole così di attivare la trascrizione. Quando il galattosio è presente, si ritiene che esso, o un suo metabolita, elimini GAL80, in modo che GAL4 sia libera di attivare i geni che codificano per gli enzimi che degradano il galattosio.

Abbiamo così appreso che questa proteina ibrida, con una superficie di attivazione costituita da soli 15 amminoacidi, potrebbe attivare geni del lievito con un'efficienza pari a circa il 20 per cento di quella della proteina GAL4 intatta. Se, invece, l'ordine degli amminoacidi venisse rimescolato, il frammento proteico non avrebbe alcuna funzione attivante, il che dimostra che la carica elettrica da sola non conferisce la capacità di attivare. Forse la particolare distribuzione delle cariche negative lungo una delle superfici dell'elica è davvero necessaria per creare una zona di attivazione.

Gli esperimenti appena descritti hanno creato le condizioni appropriate per un'ulteriore generalizzazione che molti di noi non si aspettavano: GAL4 e i vari frammenti derivati da essa, che attivano la trascrizione nel lievito, svolgono la stessa funzione anche nelle cellule dei mammiferi, in quelle del moscerino della frutta (*Drosophila*) e perfino in quelle di una pianta, il tabacco. Gli esperimenti grazie ai quali è stato possibile dimostrare questi fatti hanno una base comune. In ogni caso, il gene per GAL4 o uno dei suoi derivati viene introdotto in una cellula estranea, che sintetizza la proteina. Si inserisce, quindi, un sito di legame per GAL4 vicino a uno dei geni originari della cellula. In ogni caso, cioè sia nelle cellule dei mammiferi sia in quelle del moscerino della frutta o del tabacco, l'attivatore del lievito incrementa, anche notevolmente, la trascrizione del gene originario.

Non si osserva alcuna attivazione se il sito di legame di GAL4 viene omesso oppure se la regione di GAL4 che si lega al DNA si esprime senza una regione attivante. Perciò una molecola che contenga una regione attivante attaccata a un frammento che si lega al DNA attiverà l'espressione genica in molti eucarioti, forse addirittura in tutti.

Dato che una regione attivante conserva la propria funzione anche quando è introdotta in cellule estranee, è possibile che essa conservi in queste cellule anche la capacità di agire a notevole distanza? È noto che nei geni degli organismi superiori prevalgono siti di legame remoti, i potenziatori. I nostri attivatori possono avviare la trascrizione anche quando si legano al DNA a una distanza superiore a 1000 coppie di basi dal gene? La risposta è positiva, ammesso che l'attivatore abbia una regione attivante particolarmente potente. Una regione di questo genere si trova in una proteina regolatrice, la vp16, sintetizzata dal virus dell'*Herpes simplex*. Quando la parte acida della proteina si fonde con il frammento di GAL4 che si lega al DNA e l'ibrido viene introdotto in cellule di mammifero, esso attiva la trascrizione di un gene che è situato più di 1000 coppie di basi a monte o a valle rispetto alla UAS_G.

Riteniamo che questo segmento di vp16 abbia proprio la giusta combinazione di struttura e carica elettrica per poter interagire in modo particolarmente efficace con la proteina bersaglio. La stessa GAL4 opera solo a distanze più piccole sul DNA, presumibilmente perché la sua regione attivante interagisce in maniera un po' meno energica con la proteina bersaglio.

Questa mia rassegna sulla regolazione genica ha fin qui trascurato le modalità attraverso le quali le proteine regolatrici sono controllate all'interno di una cellula in modo da poter mediare

una risposta adeguata alle condizioni ambientali. Sappiamo, almeno parzialmente, quanto accade per GAL4. Poiché questa proteina attiva i geni che degradano il galattosio, è importante che essa operi quando questo zucchero è presente nell'ambiente in cui si trova la cellula di lievito e non quando esso è assente. Si è visto che una proteina inibitrice, la GAL80, fa ordinariamente da schermo alle superfici attivanti di GAL4; quando il galattosio è presente, questo zucchero o uno dei suoi prodotti metabolici toglie di mezzo l'inibitore, portando allo scoperto le superfici di attivazione.

Gli studi e i risultati che ho presentato in questo articolo mettono in evidenza anche un buon numero di problemi irrisolti. In primo luogo, se ammettiamo che le regioni attivanti interagiscano con qualche altra proteina interessata nella trascrizione, qual è allora quest'altra proteina? Potrebbe trattarsi della stessa RNA-polimerasi, ma sospettiamo piuttosto che sia una proteina diversa che agisce da intermediario tra la proteina regolatrice e la polimerasi. Evidentemente questa proteina bersaglio è presente in una forma simile in molti eucarioti (forse in tutti).

In secondo luogo, qual è la natura dell'interazione tra una regione attivante e il suo bersaglio? In biologia molecolare si è abituati a pensare alle interazioni proteina-proteina come se fossero determinate dalle specifiche strutture delle molecole interagenti, ma in questo caso molecole che hanno una varietà di sequenze con cariche negative operano con buona efficienza.

Infine, è importante notare che gli attivatori che abbiamo studiato possono rappresentare soltanto una classe di proteine attivatrici; potrebbero esservene altre.

BIBLIOGRAFIA

DUNN TERESA M., HAHN STEVEN, OGDEN SHARON e SCHLEIF ROBERT F., *An Operator at −280 Base Pairs That Is Required for Repression of araBAD Operon Promoter: Addition of DNA Helical Turns between the Operator and Promoter Cyclically Hinders Repression* in «Proceedings of the National Academy of Sciences of the United States of America», 81, n. 16, agosto 1984.

HOPE IAN A. e STRUHL KEVIN, *Functional Dissection of a Eukaryotic Transcriptional Activator Protein, GCN4 of Yeast* in «Cell», 46, n. 6, 12 settembre 1986.

PTASHNE MARK, *A Genetic Switch: Gene Control and Phage λ*, Cell Press & Blackwell Scientific Publications, 1987.

PTASHNE MARK, *How Eukaryotic Transcriptional Activators Work* in «Nature», 335, n. 6192, 20 ottobre 1988.

Il mescolamento dei fluidi

Esperimenti e modelli al calcolatore consentono di spiegare come fluidi viscosi che scorrono in due dimensioni con moto periodico possano generare il caos necessario a un efficiente mescolamento

di Julio M. Ottino

Che cosa hanno in comune l'eruzione del Krakatoa, la preparazione della pasta sfoglia e la luminosità delle stelle? Tutti e tre hanno in comune una qualche forma di mescolamento. Un violento mescolamento di magmi potrebbe avere innescato l'eruzione del Krakatoa; spianando e ripiegando l'impasto (un tipico processo di mescolamento) si ottiene la sfoglia, mentre è il mescolamento che avviene all'interno di una stella a determinarne la composizione chimica e quindi la luminosità superficiale. Esempi di mescolamento si possono trovare letteralmente in tutto l'universo in un grandissimo intervallo di scale di tempi e di dimensioni. I gas emessi si mescolano con l'aria dell'ambiente in pochi secondi, mentre i processi di mescolamento che hanno luogo nel mantello terrestre possono durare anche parecchie centinaia di milioni di anni.

Il mescolamento ha un ruolo critico anche nella moderna tecnologia. Gli ingegneri chimici si basano sul mescolamento per assicurarsi che le sostanze reagiscano adeguatamente per produrre miscele di polimeri dotate di proprietà particolari e per disperdere nelle condutture agenti che riducano la resistenza. Tuttavia, nonostante la sua ubiquità in natura e nell'industria, il mescolamento è conosciuto in modo incompleto. In effetti, i ricercatori non possono neppure basarsi su una terminologia comune: il mescolamento è spesso chiamato «rimescolamento» in oceanografia e in geofisica, «miscelazione» in ingegneria chimica e «agitazione» nell'ingegneria dei processi produttivi.

Indipendentemente dal nome, non vi è dubbio che il mescolamento sia un processo di estrema complessità e sia presente in una grande varietà di sistemi. Per esempio, per costruire una teoria del mescolamento dei fluidi si devono prendere in considerazione fluidi che possono essere miscibili o parzialmente miscibili, reagenti o inerti, e flussi che sono lenti e ordinati oppure molto veloci e turbolenti. Non sorprende perciò che una teoria unica non sia in grado di spiegare tutti gli aspetti del mescolamento dei fluidi e che semplici calcoli di routine non riescano solitamente a catturare tutti i particolari importanti.

Eppure, sia esperimenti fisici, sia simulazioni al calcolatore possono fornire informazioni sul processo di mescolamento. Negli ultimi anni i miei colleghi e io abbiamo seguito entrambi i metodi nel tentativo di migliorare la comprensione di vari aspetti del processo, in particolare del mescolamento relativo a flussi lenti e a fluidi viscosi quali gli oli.

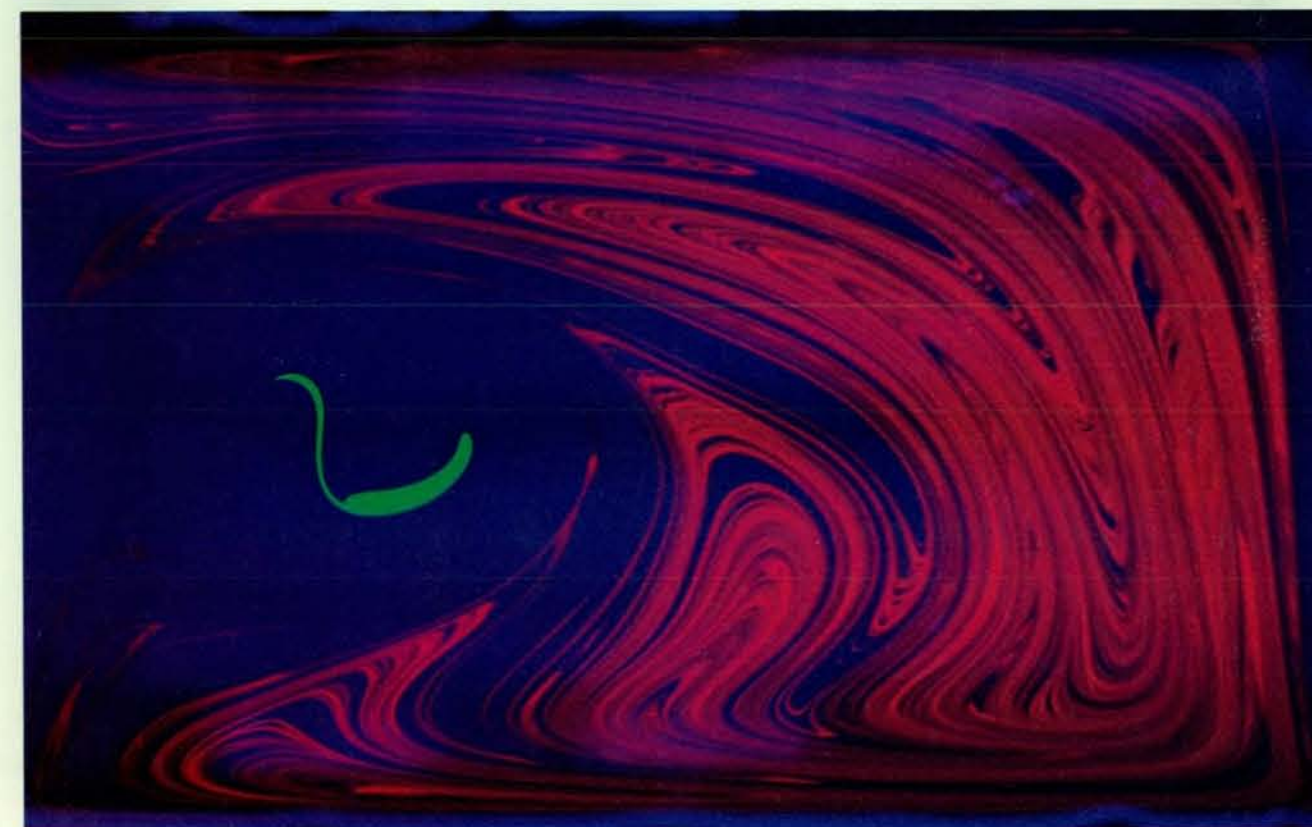
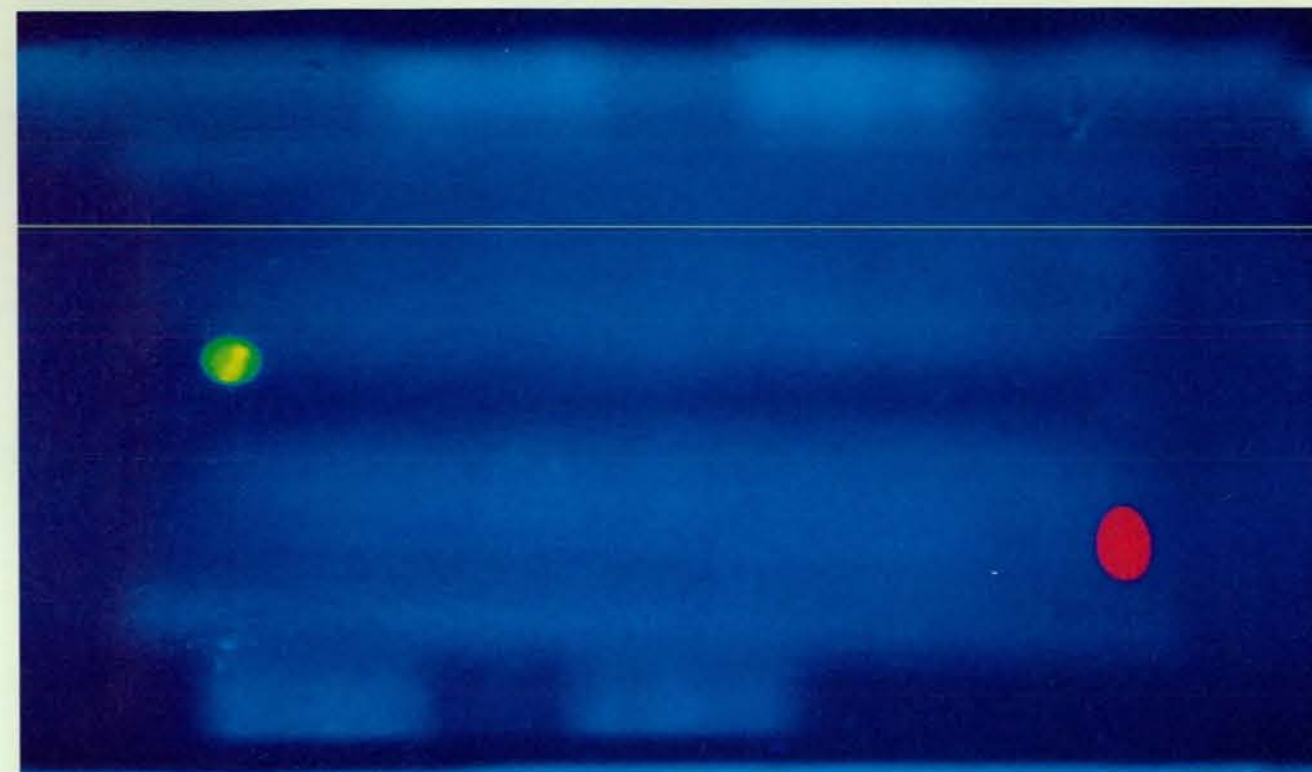
Unire insieme due colori a olio è un buon esempio di mescolamento di fluidi viscosi. Dopo solo pochi secondi di mescolamento si può produrre una sconcertante figura di striature allungate e ripiegate. (I rilegatori sfruttano il fenomeno nella «marmorizzazione» che orna talvolta le copertine o i risguardi dei libri.) Tuttavia, a meno che non si sia voluto intenzionalmente il contrario, è possibile trovare alcune «isole» di colore non mescolato fra striature a forma di spirale. Anche se il mescolamento di fluidi viscosi può produrre strutture fantasticamente complesse, esso può anche produrre figure con un certo grado di regolarità e coerenza.

I miei studenti e io all'Università del Massachusetts ad Amherst abbiamo cercato di caratterizzare i flussi che producono tali figure eseguendo esperimenti e simulazioni al calcolatore che si rifanno al mescolamento di due colori. In alcuni dei nostri esperimenti iniettiamo «gocce» di glicerina colorata in una massa di glicerina incolore contenuta in una cavità profonda. Quando le pareti della cavità vengono fatte muovere periodicamente, le forze di taglio che esse esercitano sul fluido viscoso contenuto nella cavità possono allungare e ripiegare la goccia colorata in modo piuttosto complesso; l'intera cavità mostra presto una complessa figura di pieghe su pieghe. Contemporaneamente però può accadere che nello stesso contenitore una goccia analoga non subisca alcun allungamento; può darsi che la goccia si muova e ruoti, ma poi torni regolarmente alla posizione iniziale. Perché si producono figure tanto diverse?

Fondamenti di meccanica dei fluidi

La chiave per comprendere gli aspetti fondamentali di questo tipo di mescolamento sta nel concetto di «moto», un'idea che si può far risalire al matematico svizzero del XVIII secolo Leonhard Euler. Il moto di un fluido è un'espressione matematica che stabilisce dove si troverà in qualsiasi istante futuro ogni particella del fluido. Se potessimo conoscere il moto di un determinato flusso, si potrebbe in linea di principio conoscere quasi tutto ciò che c'è da sapere sul mescolamento che esso produce. Per esempio, si potrebbero calcolare le forze e l'energia totale necessarie per ottenere un certo grado di mescolamento nel sistema.

Durante il secolo scorso la descrizione del flusso in termini di moto di un fluido è stata in gran parte abbandonata a favore di una descrizione basata sul campo delle velocità del fluido: un'espressione che specifica la velocità del fluido in ogni punto della regione di flusso in un istante qualsiasi. Però, se si conosce il moto, è facile calcolare il campo delle velocità, mentre la conoscenza del campo delle velocità non è sufficiente per calcolare esplicitamente il moto. Essendo quindi il moto in un certo senso una descrizione maggiormente basilare del flusso, i miei collaboratori e io abbiamo preferito lavorare con quello che forse molti considerano un concetto superato.

Strettamente connesso al concetto di moto è quello di trasformazione puntuale, un'operazione matematica che consente di identificare una particella di fluido e di specificarne la posizione in ogni istante futuro. Ogni particella di fluido viene «mappata» in una nuova posizione

Sia i flussi caotici sia quelli non caotici sono evidenti in un esperimento eseguito da Kenny Leong e dall'autore nel loro laboratorio all'Università del Massachusetts ad Amherst. Una cavità rettangolare viene riempita di glicerina e due «gocce» di tracciante fluorescente (una verde e una rossa) sono iniettate al di sotto della superficie (*in alto*). Ogni parete della cavità può scorrere parallelamente a se stessa indipendentemente dalle altre. In questo esperimento le pareti superiore e inferiore vengono fatte muovere periodicamente, ma in modo discontinuo. La parete superiore si sposta da sinistra verso destra per un certo tempo, quindi si ferma; a questo punto la parete inferiore si sposta alla stessa velocità e per lo stesso intervallo di tempo, ma da destra verso sinistra. I due spostamenti costituiscono un periodo. Dopo 10 periodi (*in basso*) la goccia rossa, situata in una regione di mescolamento caotico, è stata allungata e ripiegata varie volte. La goccia verde è stata allungata solo in misura ridotta: essa evidenzia un'«isola» di mescolamento non caotico.

applicando la trasformazione. Particelle identificate inizialmente come separate non possono occupare la stessa posizione nello stesso istante e una particella non può dividersi in due. Anche se una trasformazione puntuale esiste in teoria per tutti i flussi di mescolamento, essa si può ottenere esattamente solo per i casi più semplici. Per questo motivo, la maggior parte di ciò che si sa sul mescolamento è limitata a flussi relativamente semplici, come i flussi lineari nei quali le linee di un tracciante non si piegano. Eppure è probabile che questi tipi di flusso non riescano a far individuare i processi che portano a un efficiente mescolamento, i quali sono per loro natura non lineari. Per avere almeno un'idea di ciò che entra in gioco in tali processi, si devono considerare flussi stazionari in due dimensioni.

Flussi bidimensionali

Tutti i flussi bidimensionali sono costituiti dagli stessi elementi: punti iperbolici (chiamati anche di sella) e punti ellittici (*si veda l'illustrazione a pagina 39*). Nel caso di punto iperbolico il fluido si muove verso di esso lungo una direzione e se ne allontana lungo un'altra. Invece nel caso di punto ellittico il fluido circola attorno al punto. (Dovrei anche ricordare che esiste un terzo tipo di punto, il punto parabolico, nel quale il moto del fluido è di taglio, ossia tangenziale. Tali punti si trovano, per esempio, in un fluido che scorre lungo una parete solida. Si possono trascurare i punti parabolici nella descrizione della natura del mescolamento nei flussi bidimensionali.) Come è prevedibile, il mescolamento in un flusso stazionario bidimensionale è piuttosto inefficiente in confronto al mescolamento in flussi tridimensionali, in particolare in quelli che cambiano continuamente nel tempo. Infatti, due sono le possibilità in un flusso stazionario bidimensionale in una regione limitata: o le particelle di fluido seguono ripetutamente le stesse traiettorie, le cosiddette linee di flusso, oppure non si muovono affatto.

Dal momento che nei flussi stazionari le linee di flusso sono fisse e le traiettorie delle particelle di fluido non possono mai incrociarsi, le particelle di fluido non hanno alcuna possibilità di venire reciprocamente in contatto, ossia di mescolarsi. Esiste un modo per liberarsi del confinamento delle linee di flusso cosicché le particelle di fluido possano evitare di dover seguire ripetutamente la stessa linea di flusso? La risposta è positiva, se si può modificare nel tempo il moto del fluido in modo che una linea di flusso a un istante intersechi quella relativa a un istante successivo.

Il modo più semplice per farlo (e il più facile per analizzarlo) è di costringere il flusso a variare periodicamente nel tempo. Perché un tale flusso porti a un mescolamento efficiente, però, esso deve essere in grado di estendere e ripiegare una regione di fluido e di riportarla - allungata e ripiegata - alla sua posizione iniziale. L'operazione di estensione e ripiegamento corrisponde a quella che viene chiamata mappa a ferro di cavallo, descritta da Stephen Smale dell'Università della California a Berkeley.

Il fatto che per poter mescolare efficientemente una sostanza si debba riportarne una parte nella sua posizione iniziale è contro ogni intuizione. Eppure, se il mescolamento avviene in un sistema limitato non vi è realmente alcuna alternativa. Se si lancia ripetutamente una freccia su un bersaglio, qualche lancio finirà con il cadere arbitrariamente vicino a qualche altro, dal momento che la superficie del bersaglio è semplicemente finita. Per lo stesso motivo, la ripetizione di allungamenti e di ripiegamenti in una cavità chiusa porterà immancabilmente, in certi istanti, particelle di fluido vicinissime alle loro posizioni iniziali.

Se in un flusso periodico una particella di fluido ritorna dopo un certo tempo nella sua esatta posizione iniziale, la particella definisce quello che viene detto un punto periodico. A seconda del numero di periodi necessari perché la particella ritorni alla propria posizione di partenza, si parla di punto periodico di periodo uno, di periodo due, e così via. Un punto periodico si può anche classificare come iperbolico o ellittico a seconda della direzione del flusso nelle sue immediate vicinanze.

Quando un punto periodico ellittico descrive la sua traiettoria ciclica, il materiale circostante non solo circola attorno a esso (come farebbe attorno a un punto ellittico fisso), ma si muove anche con esso. Tuttavia, nonostante la rotazione e la traslazione del materiale, questo non cede facilmente materia nella restante parte del flusso. Tali regioni di materiale sono viste come «isole» di fluido e il mescolamento all'interno delle isole è tipicamente lento. Dato che il materiale non può né entrare né abbandonare le vicinanze di un punto periodico ellittico, tali punti costituiscono ostacoli a un efficiente mescolamento.

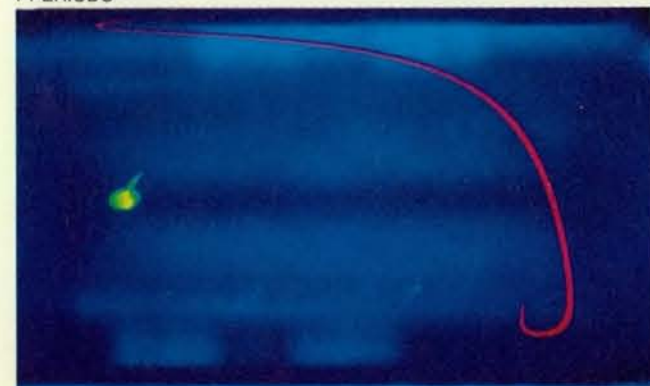
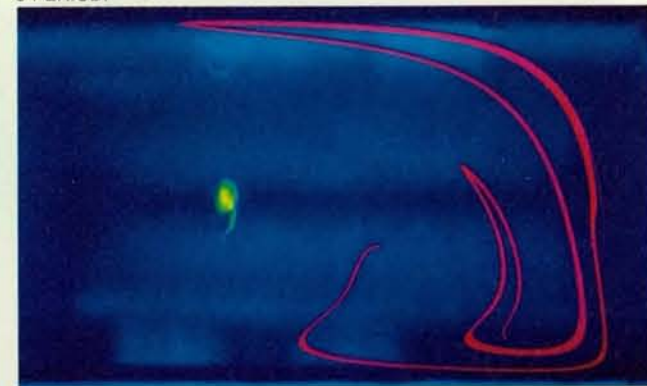
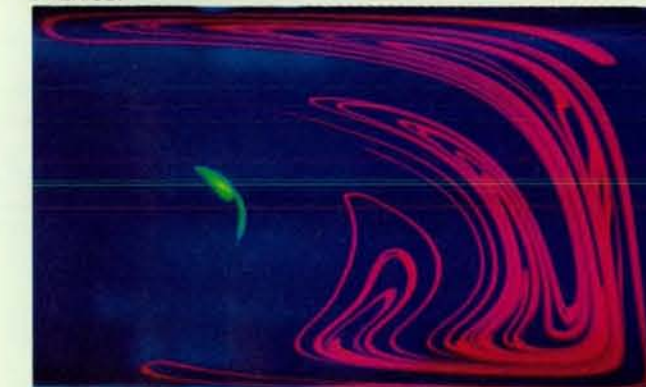
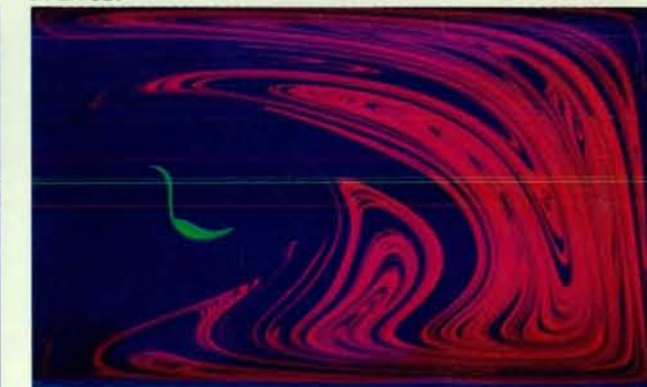
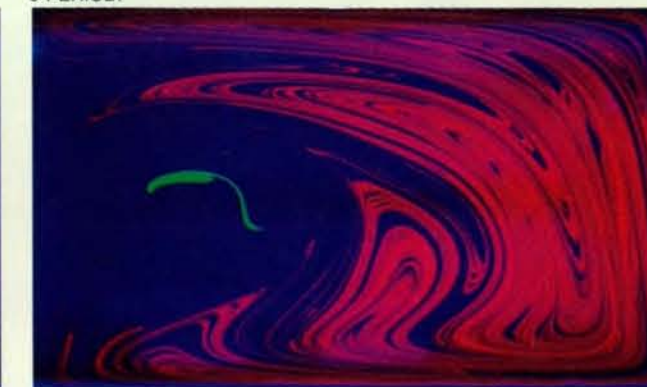
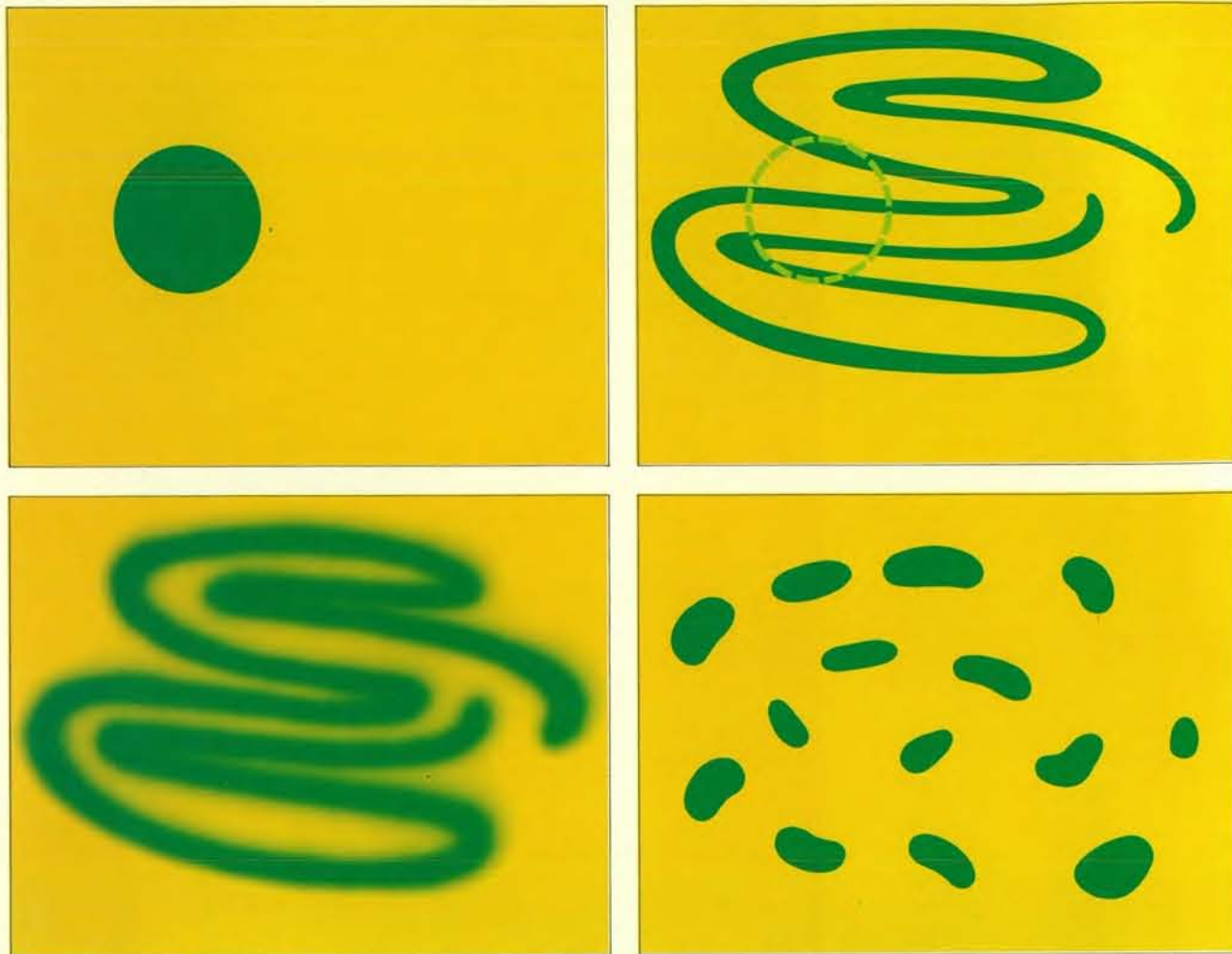
Del pari, mentre un punto periodico iperbolico descrive la sua traiettoria ciclica, il materiale circostante che si muove con il punto subisce una contrazione in una direzione e un allungamento in un'altra. Nel fare ciò il punto spande filamenti allungati di fluido in una direzione e attira materiale in un'altra. (Se supponiamo che i fluidi siano incomprimibili, l'allungamento e la contrazione devono compensarsi a vicenda.)

Le impronte digitali del caos

Dove va a finire il materiale sparso da un punto periodico iperbolico? Da dove viene il materiale che si avvicina al punto? Può darsi che un afflusso si fonda in modo regolare con un efflusso, cioè che il materiale sparso da un punto iperbolico venga attirato dallo stesso o da un altro punto iperbolico. Questo è proprio ciò che accade in un flusso stazionario (sebbene in tal caso i punti iperbolici siano fissi e non periodici), con il risultato che il flusso non deforma e non ripiega il materiale in maniera efficiente.

I flussi bidimensionali dipendenti dal tempo possono però tradursi in efficienti meccanismi di allungamento e di ripiegamento poiché in tali flussi è possibile che una regione di efflusso associata a un punto periodico iperbolico attraversi la regione di afflusso dello stesso o di un altro punto iperbolico. Un punto nel quale l'afflusso e l'efflusso di un solo punto iperbolico si intersecano viene detto «punto omoclino trasversale». Se l'intersezione deriva da flussi di due punti iperbolici diversi, esso viene detto «punto eteroclino trasversale».

Le intersezioni omocline ed eterocline sono le impronte digitali del caos. Da un punto di vista matematico, allora, un sistema in grado di produrre mappe a ferro di cavallo o intersezioni trasversali omocline o eterocline si può classificare come caotico. Ne consegue che una mappa a ferro di cavallo implica in realtà l'esistenza di punti trasversali omoclini; analogamente l'identificazione di uno solo di tali punti è sufficiente a implicare l'esistenza di una mappa a ferro di cavallo.

Il fatto che un solo attraversamento di afflusso e di efflusso conduca invariabilmente a punti trasversali omoclini, e che un siffatto attraversamento possa aver luogo perfino in quelli che paiono sistemi fisici «ben educati» descritti dalle leggi del moto di Newton fu scoperto per la prima volta dal matematico francese del XIX secolo Henri Poincaré. Tuttavia l'analisi del comportamento straordinariamente complesso derivante da tali attraversamenti (oggi chiamato «caos») mise

L'allungamento e il ripiegamento tipici del mescolamento caotico sono evidenziati dalla goccia rossa in questa sequenza di fotografie dell'esperimento mostrato nell'illustrazione della pagina precedente. Dopo tre periodi è chiaramente visibile lo schema fondamentale di allungamento e ripiegamento. L'isola verde che contraddistingue una regione di mescolamento essenzialmente non caotico e le pieghe che caratterizzano una regione di mescolamento caotico si muovono nella cavità, ma dopo ogni periodo tornano nelle stesse posizioni (anche se alquanto deformate). I filamenti tracciati dalla goccia verde indicano che essa subisce una rotazione completa ogni due periodi circa. Se si eseguisse l'esperimento all'indietro, la goccia verde ritornerebbe quasi alla forma e alla posizione iniziali, dato che l'errore nella riproduzione dei suoi movimenti in senso inverso cresce linearmente. È invece praticamente impossibile lo «smescolamento» della goccia rossa: in tal caso l'errore nel riprodurre i suoi movimenti in senso inverso cresce esponenzialmente.

Il mescolamento di fluidi nei processi naturali o industriali coinvolge, oltre all'allungamento e al ripiegamento, la diffusione molecolare e la rottura. Un goccia di vernice colorata in una massa fluida (*in alto a sinistra*) può essere allungata e ripiegata indefinitamente senza diffondere o rompersi (*in alto a destra*) solo in un caso ideale. In tale situazione, per ottenere un mescolamento efficiente, una parte della goccia deve tornare alla posizione iniziale. La diffusione molecolare (senza cui è impossibile il totale mescolamento) rende normalmente indistinguibili i confini tra i fluidi miscibili (*in basso a sinistra*). Nel caso di fluidi immiscibili la goccia, nell'allungarsi, può spezzarsi in goccioline che possono successivamente aderire per formare molte gocce più piccole di quella iniziale (*in basso a destra*).

in crisi Poincaré, il quale decise di non esplorare ulteriormente l'argomento.

Nella stessa misura in cui il mescolamento può essere rappresentato da una trasformazione puntuale deterministica, esso dovrebbe essere cinematicamente reversibile. In altre parole, dovrebbe essere possibile «riseparare» i fluidi (perlomeno se si trascura la diffusione molecolare). Tuttavia, l'esperienza quotidiana suggerisce che il mescolamento è un processo irreversibile. Anche se il sistema è in teoria deterministico, i moti che conducono a ripetuti allungamenti e ripiegamenti non si possono invertire.

Una situazione abbastanza simile esiste in altri sistemi fisici, quali quelli analizzati da Poincaré, formati da molte particelle i cui moti sono descritti da equazioni deterministiche. (Questi tipi di sistemi vengono comunemente detti «sistemi hamiltoniani».) Uno dei più famosi fisici americani del XIX secolo, J. Willard Gibbs, scoprì che persino i sistemi hamiltoniani possono avere una irreversibilità e una imprevedibilità intrinseche, e il fatto che egli sia ricorso per spiegarlo a un esperimento ideale relativo al mescolamento dà la misura della sua intuizione. Le sue considerazioni passarono apparentemente inosservate fino a quando l'oceanografo svedese Pierre Wellander non le sottolineò in un acuto articolo pubblicato nel 1955.

Come catturare il caos in un flusso

Il fatto che l'allungamento e il ripiegamento abbiano un ruolo di primo piano nel mescolamento era noto in ingegneria chimica fin dagli anni cinquanta in seguito alla ricerca pionieristica di Robert S. Spencer e Ralph M. Wiley della Dow Chemical Company e di William D. Mohr e collaboratori della E. I. du Pont de Nemours & Company, Inc., mentre le conseguenze dell'esistenza di mappe a ferro di cavallo e di punti omoclini ed eteroclini sono state riconosciute solo recentemente.

Pare che sia stato il matematico russo Vladimir I. Arnold ad aver trovato la prima correlazione diretta tra caos e flussi di fluidi. Secondo Michel Hénon, un astronomo francese dell'Osservatorio di Nizza, Arnold suggerì nel 1965 la possibilità che sistemi fluido-meccanici possano mostrare traiettorie caotiche di particelle. Hénon approfondì l'ipotesi di Arnold e in un articolo di tre pagine che conteneva una sola illustrazione fu capace di dimostrare che un flusso stazionario tridimensionale di un fluido non viscoso può dare effettivamente origine a linee di flusso caotiche.

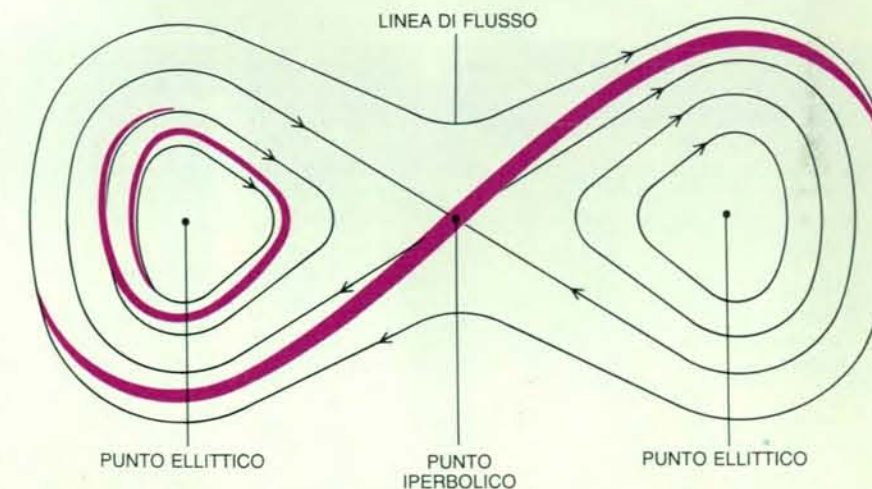
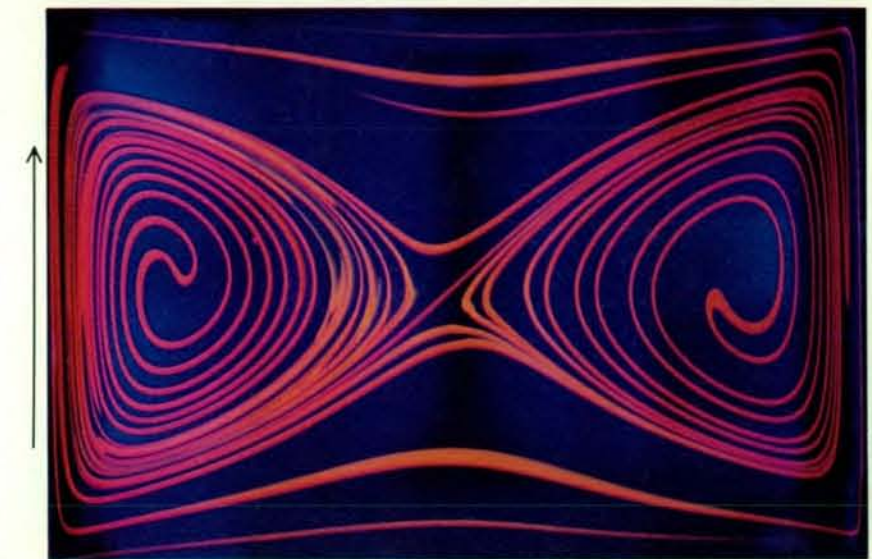
Nel 1984 Hassan Aref, allora alla Brown University, osservò che le equazioni che descrivono le traiettorie delle particelle di fluido in un flusso bidimensionale sono formalmente identiche a quelle che descrivono un sistema hamiltoniano. Egli spinse oltre le sue osservazioni, dimostrando, mediante una simulazione al calcolatore, che un sistema hamiltoniano soggetto a forze periodiche può in realtà produrre un mescolamento efficiente.

A tre dimensioni l'analogia tra mescolamento e sistemi hamiltoniani non funziona, mentre a due dimensioni l'analogia è esatta: il mescolamento di fluidi si può considerare come una rappresentazione visiva del comportamento di un sistema hamiltoniano caotico. La ricerca di Aref, tenendo conto del fatto che in laboratorio è molto più facile studiare flussi bidimensionali che non flussi tridimensionali, mi ha suggerito di cercare tracce di caos in un sistema sperimentale di flusso in una cavità che insieme ai miei studenti ho costruito ad Amherst nel 1983.

Immagini di flussi

Kenny Leong, un mio studente diplomato, e io riuscimmo a determinare le posizioni approssimative di alcuni punti periodici e di strutture di grande scala in flussi bidimensionali registrando accuratamente immagini stroboscopiche del sistema in movimento. (Essendo noi interessati al mescolamento rapido, ci siamo concentrati sul comportamento di punti periodici di ordine basso, cioè con periodo uno, due o tre; i punti di ordine superiore non partecipano al processo con la stessa frequenza di quelli di ordine più basso.) In un tipico esperimento introduciamo gocce di vernice fluorescente in certe posizioni della cavità rettangolare, illuminiamo la cavità con luce ultravioletta, mettiamo in movimento i suoi lati secondo un particolare schema di moto e registriamo le posizioni delle gocce e le relative contorsioni riprendendo fotografie del sistema a intervalli regolari. Se il mescolamento è efficiente, le particelle di vernice esplorano una grande regione del sistema. Se invece il mescolamento è modesto, le gocce cedono il colorante solo lentamente alla massa del fluido oppure rimangono vicine a punti periodici ellittici.

In altri esperimenti Paul D. Swanson, un altro dei miei studenti, e io ci siamo occupati specificamente di flussi che trovano una soluzione analitica esatta delle equazioni di moto del fluido. In tal modo siamo in grado di confrontare al meglio i nostri risultati sperimentali con quelli previsti dalla teoria. Sfortunatamente, il numero di sistemi per i quali esistono soluzioni analitiche esatte è piuttosto piccolo, e molti sono talmente idealizzati da non poter essere riprodotti in esperimenti di laboratorio. Uno dei sistemi che ammette soluzione esatta ed è realizzabile sperimentalmente è il flusso tra due cilindri eccentrici rotanti. Un siffatto sistema è stato studiato anche da Aref (oggi all'Università della California a San Diego) e da Michael Tabor e Rene Chevray della Columbia University.

Esperimenti a tappeto su flussi caotici bidimensionali rivelano che le strutture fluide di mescolamento di grande scala (quali la posizione e la forma di isole e di grandi pieghe) sono perfettamente riproducibili; i più piccoli particolari delle strutture allungate e ripiegate non lo sono. Il motivo dipende dal fatto che piccole deviazioni dalla posizione iniziale delle gocce di vernice vengono amplificate all'interno delle regioni caotiche di flusso. Ed è proprio così che deve essere: dovrebbe essere impossibile ripro-

I punti ellittici e iperbolici sono aspetti tipici di flussi lenti bidimensionali. La fotografia (*in alto*), eseguita da Leong e dall'autore, mostra uno di tali flussi, prodotto quando le pareti opposte di una cavità rettangolare riempita di glicerina vengono spostate in senso opposto con velocità costante. Le linee in arancione (prodotte da un tracciante iniettato all'inizio lungo una linea che va dall'angolo a sinistra in basso a quello a destra in alto) sono quasi allineate con le linee di corrente del flusso, cioè le linee descritte nel flusso stazionario dalle particelle di fluido in moto. L'andamento di flusso contiene tre punti fissi: un punto centrale iperbolico e due punti ellittici ai lati. Il flusso nelle vicinanze di ogni punto ellittico (*in basso*) produce un vortice che ruota in senso orario; esso fa aumentare linearmente con il tempo la lunghezza del tracciante. Il flusso nelle vicinanze di un punto iperbolico si avvicina al punto in una direzione e se ne allontana in un'altra. Poiché il materiale del fluido non può attraversare le linee di flusso, un flusso stazionario bidimensionale non è efficiente nel mescolamento. Tuttavia, se si fa variare il flusso nel tempo, i filamenti allungati del tracciante non hanno sufficiente tempo per allinearsi con nuove linee di flusso e vengono rapidamente ripiegati da un cambiamento di direzione del flusso.

durre esattamente qualsiasi fase dei nostri esperimenti di mescolamento. In fondo, obiettivo del mescolamento è creare una distribuzione casuale. Ed è esattamente ciò che si ottiene con i meccanismi di allungamento e di ripiegamento impiegati negli esperimenti.

È anche interessante notare come il caos possa pacificamente coesistere con simmetrie nel flusso, quali quelle rappresentate da punti periodici. Infatti, eliminando sistematicamente le simmetrie in un flusso caotico, i miei collaboratori e io siamo riusciti ad aumentare l'efficienza di mescolamento del flusso.

Sperimentazione ed elaborazione a confronto

Se il sistema sperimentale è piuttosto semplice (tale da potersi ricavare un'espressione matematica del campo delle velocità), esso può essere facilmente simulato con un calcolatore. In un programma tipico viene strategicamente disposto un certo numero di «particelle» di prova in un moto simulato, ossia in un campo delle velocità; le posizioni calcolate delle particelle dopo circa 1000 periodi offrono in tal caso un buon quadro del comportamento generale del sistema dopo che è stato messo in funzione per molto tempo. L'immagine ottenuta con questo tipo di simulazione è detta «sezione di Poincaré» e una sezione di Poincaré di aspetto complesso viene spesso assunta come prova al calcolatore dell'esistenza del caos (*si veda la parte superiore dell'illustrazione della pagina a fronte*). Le simulazioni del mescolamento al calcolatore presentano anche una forma di irreversibilità cinematica, ma nel loro caso essa deriva dall'ingrandimento esponenziale degli errori introdotti dal calcolatore, che può operare soltanto con numeri formati da un numero finito di cifre.

Se esistono simulazioni di mescolamento al calcolatore, perché preoccuparsi di eseguire esperimenti fisici? La prima cosa da tenere presente è che la risoluzione del campo delle velocità deve essere molto più elevata per simulazioni di mescolamento che per la maggior parte degli altri problemi di meccanica dei fluidi.

Anche campi delle velocità piuttosto semplici possono dare origine a strutture estremamente complesse (*si vedano le illustrazioni alle pagine 36 e 37*); in alcuni problemi di mescolamento sarebbe interessante risolvere alcuni dei particolari più fini delle strutture.

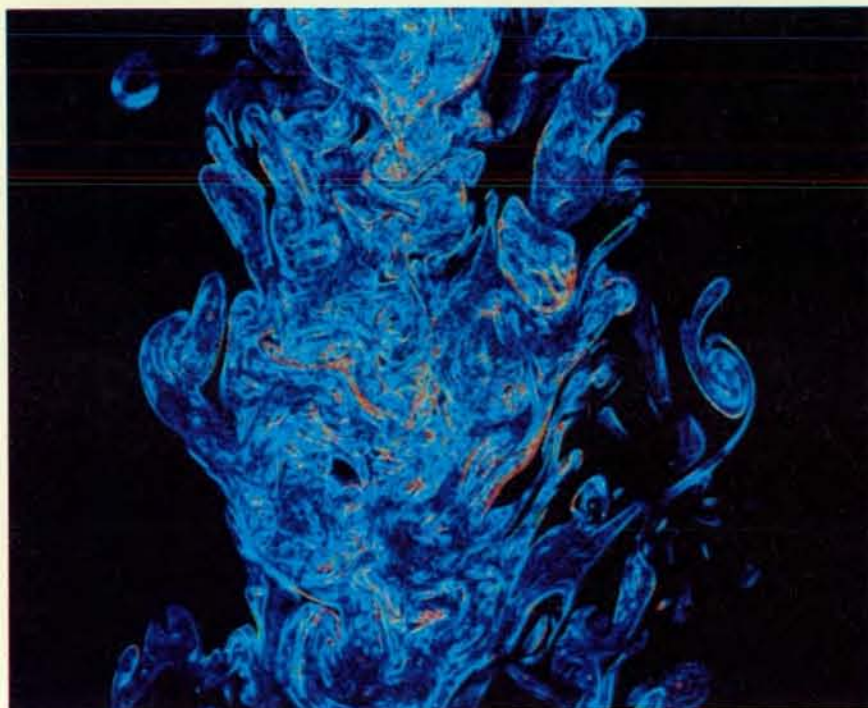
Per esempio, in una simulazione del flusso in una cavità rettangolare un campo delle velocità calcolato convenzionalmente potrebbe rivelarsi troppo grezzo per poter evidenziare i particolari delle striature allungate e ripiegate. Sarebbe anche virtualmente inutile per individuare le esatte posizioni dei punti periodici che determinano il comportamento complesso dei flussi caotici. Inoltre, mentre nella maggior parte dei problemi di meccanica dei fluidi l'obiettivo è quello di ottenere un'approssimazione del campo delle velocità, nel mescolamento il problema inizia, anziché terminare, con la specificazione del campo delle velocità.

Per tale motivo gli studi sul mescolamento sono stati in gran parte focalizzati su quelle che sono in realtà caricature di flussi (descritte da equazioni che in alcuni casi si possono risolvere esattamente) anziché su problemi più realistici (le cui soluzioni possono essere soltanto approssimate). In effetti, i metodi numerici per approssimare le soluzioni delle equazioni della meccanica dei fluidi introducono spesso effetti spuri che non esistono veramente nei problemi reali di mescolamento dei fluidi.

Comunque simulando con il calcolatore i flussi semplificati dei nostri esperimenti si incontrano spesso difficoltà insormontabili. Il calcolatore tratta un fluido come se fosse formato da elementi discreti. In una simulazione centinaia di migliaia di tali elementi possono costituire una singola goccia di vernice e il numero di calcoli necessari per tracciare una mappa del suo comportamento caotico in un flusso di mescolamento può essere enorme.

Per registrare tutte le striature in regioni di mescolamento caotico in un esempio relativamente semplice, quale quello mostrato nelle illustrazioni delle pagine 36 e 37, sarebbero necessari 300 anni di tempo macchina di un calcolatore capace di effettuare un milione di ope-

Il modello a frullatore, ideato da John G. Franjione e dall'autore, serve a illustrare il processo fondamentale di allungamento e ripiegamento caratteristico del mescolamento (*a*). Una linea tracciata sulla superficie di una cella di fluido viene allungata e ripiegata quando una lama del frullatore avanza nel fluido prima in direzione perpendicolare alla linea (*b*) e poi in direzione parallela a essa (*c*). La linea viene allungata senza spezzarsi; qualsiasi parte che si estenda oltre la sommità della cella rientra dal basso, mentre le parti che si estendono oltre la parete di sinistra rientrano dalla parete di destra. Un calcolatore può produrre immagini della cella che dipendono dal numero di volte in cui le lame hanno attraversato la cella. Nelle immagini mostrate una sola linea iniziale formata da 100 000 punti è stata allungata e ripiegata 16 volte in condizioni di mescolamento differenti. Il mescolamento che ne risulta può essere confinato a regioni della cella (*d*) o può estendersi a tutta la cella (*e*), a seconda del «vigore» con il quale le lame agitano il fluido.

Il flusso viscoso in un cuscinetto, ovvero il flusso tra due cilindri rotanti eccentrici, può essere simulato al calcolatore. Se i cilindri vengono fatti muovere periodicamente in versi opposti, il flusso conduce a un mescolamento caotico, come si vede nella sezione di Poincaré del sistema per 1000 periodi (*in alto*) e nella sua mappa di allungamento per 10 periodi (*in basso*). Una sezione di Poincaré viene prodotta inserendo un certo numero di «particelle» colorate di prova nel flusso che simula un sistema di mescolamento, calcolando il loro moto per ogni periodo e spostandole nelle loro nuove posizioni. Una mappa di allungamento indica le regioni nelle quali il fluido è stato allungato in un flusso simulato. La maggior parte dell'allungamento avviene all'interno delle regioni bianche, mentre nelle regioni colorate l'allungamento è modesto. La mappa di allungamento mostrata è sorprendentemente simile alla struttura prodotta nel flusso reale (*si veda l'illustrazione di copertina*). Le immagini sono state realizzate ad Amherst da Paul D. Swanson e dall'autore.

Il flusso turbolento può generare strutture molto diverse da quelle prodotte in un flusso viscoso lento. L'immagine, ottenuta da K. R. Sreenivasan della Yale University, è una ricostruzione al calcolatore di un getto d'acqua espulso da un ugello circolare in acqua quieta. Le strutture del flusso venivano originariamente registrate su pellicola sciogliendo un colorante fluorescente nell'acqua espulsa e dirigendo una lama di luce laser lungo l'asse dell'ugello. L'intensità della fluorescenza risultante è proporzionale al gradiente di concentrazione relativo del colorante nell'acqua; le immagini sono state codificate in colore dal blu scuro al rosso a seconda del gradiente di concentrazione. Il flusso turbolento mostrato appare formato da varie strutture frattali sovrapposte, tra cui numerosi vortici.

razioni in virgola mobile al secondo. A dire il vero, si potrebbe ritenere che non sia necessaria un'indagine così dettagliata e che sarebbe meglio una valutazione statistica dell'allungamento. Ma non sarebbe un'ammissione di incapacità? Se si può conoscere esattamente il campo delle velocità (o il moto), perché si dovrebbe affrontare il problema dal punto di vista statistico?

Per concludere, è necessario associare sviluppi teorici a esperimenti ben progettati, poiché - con molta probabilità - le tecniche di elaborazione da sole non sono in grado di rispondere a molte domande sui flussi caotici.

Per esempio, a quali tipi di moto devono essere soggette le pareti di una cavità per poter ridurre le dimensioni di tutte le isole in essa contenute (comprese quelle nuove che potrebbero comparire spontaneamente) al di sotto di un determinato livello? La risposta a questa domanda potrebbe un giorno consentire la progettazione di un sofisticato sistema di riconoscimento delle forme che sia in grado di rilevare la presenza di isole in un sistema di mescolamento e di alterare successivamente il flusso in modo tale da mescolare le isole al resto del fluido.

Limitazioni e complicazioni

Prima però di poter costruire questo tipo di macchine «intelligenti» di mescolamento, si dovrà saperne molto di più sui flussi reali. Anche se gli esperimenti e le simulazioni al calcolatore descritti in questo articolo forniscono qualche informazione su problemi generali del mescolamento (per esempio, come aumentare esponenzialmente la superficie di contatto tra due fluidi), essi rappresentano casi di problemi ideali piuttosto particolari. Per fare un esempio, i flussi in cavità descritti in questo articolo non presentano inerzia. In altre parole, il flusso cessa non appena le pareti della cavità smettono di muoversi. Di conseguenza tali flussi non portano ad alcuno dei processi caratteristici osservati nei flussi turbolenti.

Per dirlo in termini più tecnici, i numeri di Reynolds (il rapporto fra le forze inerziali e quelle viscose in un fluido) dei flussi studiati nei nostri esperimenti erano bassi. Flussi caratterizzati da bassi numeri di Reynolds (i cosiddetti flussi laminari) sono ordinati e regolari, mentre quelli caratterizzati da numeri di Reynolds elevati producono campi delle velocità variabili nel tempo piuttosto complessi che portano a un rapido mescolamento. Un osservatore in un punto fisso della nostra cavità sperimentale vedrebbe ripetersi periodicamente lo stesso semplice campo delle velocità anziché vedere i campi non periodici e imprevedibili prodotti in un flusso turbolento. Tuttavia è proprio a causa della turbolenza che è più facile mescolare la panna nel caffè con un cucchiaino (un sistema con un numero di Reynolds relativamente alto) che mescolare due vernici colorate per interni con una spatola (un sistema con un numero di Reynolds basso).

Pur avendo in un certo senso escluso dalla trattazione i flussi di mescolamento più efficaci (quelli turbolenti), vi è motivo di credere che alcune idee presentate in questo articolo possano nondimeno portare a concetti utili per lo studio di tali flussi. Per esempio, versioni lievemente più elaborate di flussi caotici bidimensionali mostrano una velocità non periodica quando sono misurate in un punto determinato. È però chiaramente necessaria una ricerca molto più approfondita prima che si possa comprendere la turbolenza altrettanto bene di quanto oggi conosciamo i flussi laminari.

In questa trattazione ho anche semplificato l'argomento supponendo che nel mescolamento la diffusione non sia importante. Ma le cose non stanno realmente così. Per prendere in considerazione l'effetto della diffusione nel mescolamento si può ricorrere a un semplice modello che ipotizza che la velocità di diffusione tra due striature contigue di due sostanze miscibili venga controllata dalla rapidità con la quale le striature vengono «strizzate» e assottigliate, la quale dipende a sua volta dalla componente del flusso in una direzione perpendicolare alle striature. In questo modo il mescolamento ha un doppio effetto che accelera la diffusione: fa aumentare la superficie di contatto tra i fluidi, riducendo nel contempo la distanza attraverso la quale i fluidi devono diffondere e aumentando i gradienti di concentrazione. Un tale modello si può in pratica estendere fino a includere l'effetto del mescolamento su reazioni chimiche come la combustione.

Un altro processo comune che - per semplicità - ho ignorato è la rottura delle goccioline nei fluidi immiscibili, che è in realtà un fenomeno molto complesso. Vi sono due casi limite: un fluido ad alta viscosità disperso in una massa di fluido a bassa viscosità e un fluido a bassa viscosità disperso in una massa di fluido ad alta viscosità. Entrambi i casi sono difficili da analizzare, ma per ragioni differenti. Nel primo caso il fluido a bassa viscosità è sottoposto all'insieme degli sforzi di taglio dal momento che esso non può trasmettere efficacemente le sollecitazioni alle goccioline del fluido ad alta viscosità. In realtà un flusso stazionario con sforzi di taglio non può spezzare una gocciolina che ha una viscosità circa quattro volte maggiore di quella del fluido che la tiene in sospensione. Sotto questo aspetto hanno più successo i flussi con grandi allungamenti di quelli con sforzi di taglio. Però i primi non potrebbero essere particolarmente efficaci nel caso in cui le goccioline a bassa viscosità siano disperse in un fluido ad alta viscosità, poiché è necessario allungare considerevolmente le goccioline prima di romperle.

I miei collaboratori e io abbiamo eseguito studi sperimentali riguardanti il mescolamento di due fluidi di diversa viscosità. Come era prevedibile, l'entità delle rotture è molto minore all'interno delle isole che nelle regioni di caos. Del resto, potrebbe capitare che un eccessivo rimescolamento possa fare aderire le goccioline fra loro; i fluidi potrebbero talvolta separarsi a causa della loro coalescenza. Usando semplici modelli al calcolatore, siamo stati in grado di prevedere la cinetica di tale aggregazione in flussi caotici semplici. Infine - e con una chiara evidenza - resta il fatto che tutti gli esperimenti da noi finora eseguiti riguardavano flussi bidimensionali, laddove nel mondo reale ci si trova a eseguire misurazioni in tre dimensioni. Solo da poco i miei studenti e io abbiamo costruito la prima apparecchiatura capace di produrre esperimenti di mescolamento controllato in flussi tridimensionali e stiamo cominciando ora i relativi esperimenti. Vi sono molte domande di base riguardanti il mescolamento in flussi tridimensionali lenti e sfortunatamente alcune delle intuizioni che abbiamo ricavato dal nostro studio dei flussi bidimensionali non si estendono necessariamente ai flussi tridimensionali.

Le striature caratteristiche del mescolamento che avviene in un flusso viscoso sono evidenti in questa roccia magmatica della catena vulcanica Inyo della California orientale. In effetti, la roccia è il prodotto del mescolamento di due magmi differenti, uno dei quali (quello caratterizzato dalle striature più chiare) conteneva microscopiche bollicine di sostanze volatili. La diffusione attraverso tali striature del magma è molto lenta; il tempo necessario perché la diffusione cancelli striature di spessore dell'ordine di un centimetro è superiore all'età della Terra. La fotografia è di Ichiro Sugioka e Bradford Sturtevant del Caltech.

Il primo passo di un lungo viaggio

L'elenco dei problemi di mescolamento non finisce qui. Il mescolamento di fluidi viscoelastici (fluidi, quale il Silly Putty, che ritornano alla loro forma originaria dopo essere stati deformati) è un problema formidabile sul quale si sa poco, nonostante il fatto che esso emerga in modo prominente nella lavorazione di polimeri ad alto peso molecolare. Il mescolamento di fluidi delicati, non in grado di sopportare variazioni rapide degli sforzi di taglio senza essere degradati, è importante in bioingegneria. Il mescolamento di fluidi altamente viscosi dovuto a moti termici interessa i geofisici che studiano il mescolamento dei magmi nel mantello terrestre.

Nonostante la scoraggiante complessità dei processi di mescolamento, si spera che possano essere compresi e che tali conoscenze possano essere poi applicate proficuamente in laboratorio e negli impianti chimici. Inoltre, poiché semplici esperimenti fungono da analoghi del caos, essi potrebbero chiarire alcuni aspetti fondamentali di sistemi caotici in generale.

BIBLIOGRAFIA

AREF H., *Stirring by Chaotic Advection* in «Journal of Fluid Mechanics», 143, giugno 1984.

RANZ WILLIAM E., *Fluid Mechanical Mixing-Lamellar Description* in «Mixing of Liquids by Mechanical Agitation», a cura di Jaromir J. Ulbrecht e Gary K. Patterson, Gordon and Breach Science Publishers, 1985.

KHAKHAR D. V., RISING H. e OTTINO J. M., *Analysis of Chaotic Mixing in Two Model Systems* in «Journal of Fluid Mechanics», 172, pp. 419-451, novembre 1986.

OTTINO J. M., LEONG C. W., RISING H. e SWANSON P. D., *Morphological Structures Produced by Mixing in Chaotic Flows* in «Nature», 333, n. 6172, 2 giugno 1988.

L'evoluzione del camoscio appenninico

Differenze etologiche hanno promosso uno studio interdisciplinare da cui è emerso che i camosci sudoccidentali appartengono a una specie diversa da quella nordorientale, di tipo alpino, evolutasi più recentemente

di Sandro Lovari

Era la fine del secolo scorso quando il marchese genovese Giacomo Doria, presidente della Società geografica italiana, in visita al suo amico professore Decio Vinciguerra, notò una pelle di camoscio proveniente dalle montagne abruzzesi, che questi teneva in casa a mo' di tappeto. Il colore e il disegno della pelle parvero a Doria alquanto diversi da quelli tipici dei camosci alpini, a lui ben noti. Così il marchese convinse l'amico a donare lo strano «tappeto» al Museo civico di storia naturale della città di Genova dove il reperto, imbalsamato in carne, restò custodito per diverso tempo in attesa di un esame più attento da parte di qualche esperto teriologo.

Fu solo qualche anno più tardi che il camoscio venne notato da uno studioso tedesco, Oscar Neumann, di passaggio per Genova. Neumann osservò il reperto con attenzione e lo trovò talmente particolare che nel 1899 pubblicò un lavoro in cui questo camoscio dell'Appennino veniva ascritto a una specie nuova, *Rupicapra ornata*, diversa da quella tipica dell'arco alpino, *Rupicapra rupicapra*. Alcuni anni più tardi, tuttavia, Neumann ci ripensò e, forse un po' timoroso di descrivere una specie nuova sulla base di un solo campione, declassò il camoscio dell'Appennino a semplice sottospecie di quello alpino.

Trascorso un decennio, lo zoologo Giuseppe Camerano ebbe la possibilità di confrontare crani, corna e pelli di camosci provenienti dai Pirenei, dall'Appennino e dalle Alpi, concludendo che esistevano tre specie: *Rupicapra pyrenaica* (Monti Cantabrici e Pirenei), *R. ornata* (Abruzzo) e *R. rupicapra* (Alpi, altri gruppi montuosi europei e dell'Asia Minore). Le prime due erano molto affini l'una all'altra e caratterizzate morfologicamente da un mantello invernale marrone scuro con cinque ampie pezzature isabella o giallastre e da una piccola macchia frontale dello stesso colore. La terza aveva un tipico mantello invernale nero o marrone molto scuro, con tre piccole macchie biancastre.

L'esteso lavoro di Camerano passò apparentemente inosservato tra i teriologi che in seguito si occuparono della sistematica dei camosci. Fu nel 1938 che il medico chirurgo e appassionato cacciatore Marcel Couturier, di Grenoble, pubblicò un importante lavoro sulla sistematica e anatomia del camoscio, dove revisionò il genere *Rupicapra*, accettando un'unica specie (*R. rupicapra*) con nove sottospecie. A queste, nel 1971, il cecoslovacco Milič Blahout aggiunse una decima. La revisione di Couturier fu universalmente accettata fino a circa 10 anni fa in tutte le pubblicazioni sui camosci.

Nel 1976 cominciai a occuparmi del comportamento sociale del camoscio appenninico, originariamente diffuso dai Monti Sibillini all'Appennino calabro, ma oggi sopravvissuto soltanto su alcune montagne del Parco nazionale d'Abruzzo in appena qualche centinaio di capi. Con mia notevole sorpresa, emersero ben presto alcune peculiarità comportamentali di questo camoscio, assenti nel repertorio di quello alpino, e viceversa.

Ben sei moduli etologici (uno tipico delle femmine e cinque dei maschi) differivano tra i due gruppi di camosci. Cinque, quelli del repertorio maschile, erano connessi al corteggiamento. Il sesto, quello femminile, era una particolare postura assunta dalle femmine di camoscio appenninico in situazioni conflittuali, ossia quando l'animale è contemporaneamente soggetto a tendenze opposte, per esempio alla fuga e all'attacco; questa postura, invece, sembra essere assente nel repertorio del camoscio alpino.

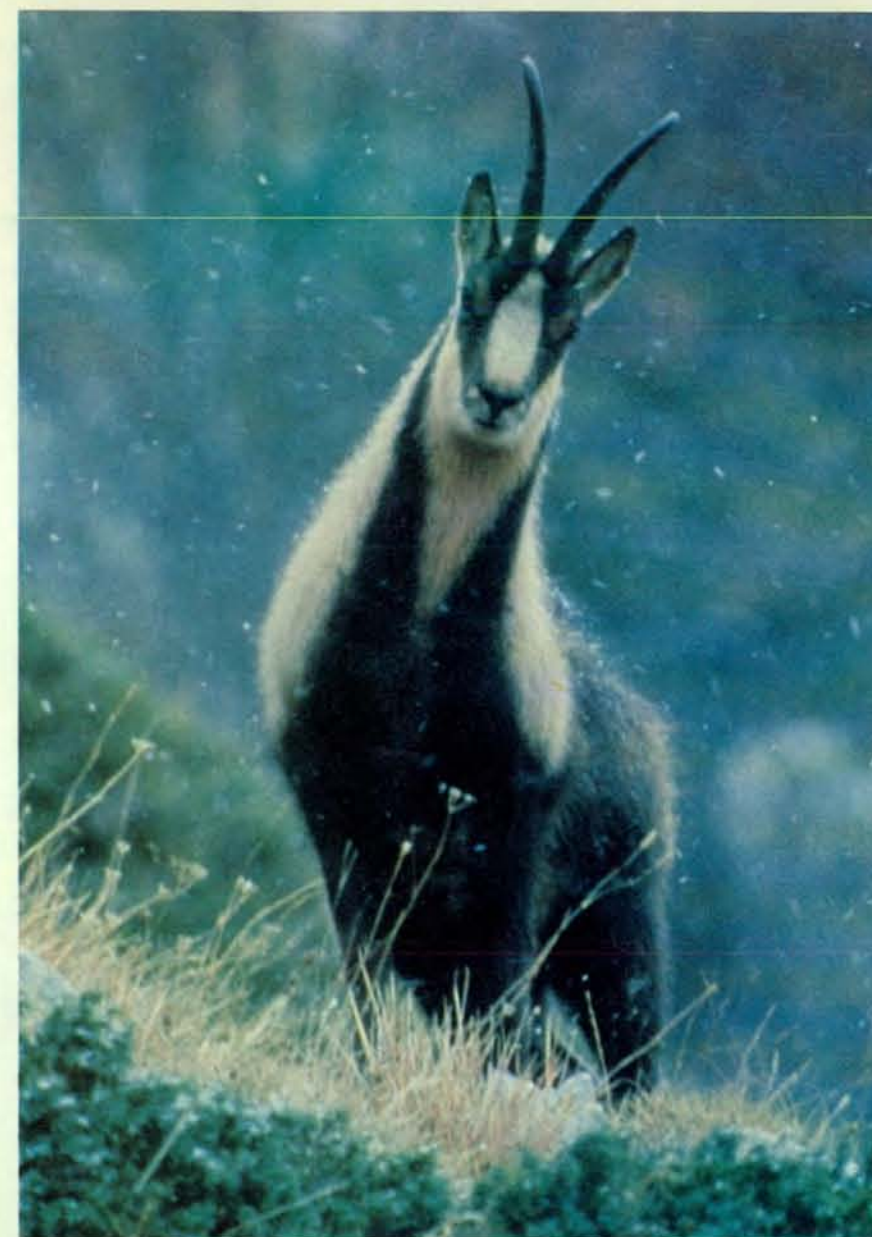
Le diversità comportamentali non sono rare da popolazione a popolazione anche nell'ambito della stessa specie, perché le contingenti caratteristiche ambientali impongono spesso forme di locale adattamento etologico agli individui che vivono e dipendono da esse. In questo caso le differenze non hanno un valore in sistematica, perché sono determinate da condizioni esterne all'individuo e non da una reale diversità intrinseca. Tuttavia, se la loro origine è genetica, queste caratterizzano invece una entità zoologica ben precisa e il loro valore in sistematica può essere grande almeno quanto ogni altro carattere ereditario di tipo morfologico, anatomico e biometrico. Sarebbe stato dunque stimolante appurare l'origine delle differenze etologiche tra il camoscio alpino e quello appenninico, dal momento che il secondo era ritenuto niente altro che una popolazione relitta del primo, vestigia della passata espansione verso sud del camoscio alpino nel corso delle ultime glaciazioni, dette del Würm (80 000-12 000 anni fa).

Le differenze comportamentali accertate tra i due tipi di camosci erano però difficilmente spiegabili in termini di diversità ambientale. Non restava pertanto che l'ipotesi di un'origine genetica.

Vennero varati quasi in parallelo studi biometrici, in collaborazione con Claudio Scala, ordinario di statistica all'Università di Siena, e genetici, in collaborazione con il genetista Luciano Bullini dell'Università di Roma «La Sapienza», prima sui soli camosci alpini e appenninici, poi coinvolgendo anche quelli dei Monti Cantabrici, dei Pirenei, della Grande Chartreuse (vicino alla valle dell'Isère in Savoia) e altre popolazioni di tipo alpino. Le differenze craniometriche statisticamente valide tra i camosci sudoccidentali (Monti Cantabrici, Pirenei, Appennino abruzzese) e quelli di tipo alpino risultarono essere almeno cinque: i primi avevano significativamente più piccole la distanza basale minima tra le corna, la lunghezza della testa e la lunghezza del cranio misurata a livello delle ossa parietali, mentre era maggiore l'angolo di impianto delle corna sulle ossa frontali. La fenestratura etmoidale, quasi sempre beante nelle popolazioni di tipo alpino, era invece sempre suturata nei camosci sudoccidentali.

A livello genetico, poi, un'analisi di 25 *loci* enzimatici eseguita su campioni di fibre muscolari appartenenti a camosci dei Monti Cantabrici, dei Pirenei, dell'Appennino abruzzese e delle Alpi occidentali, centrali e orientali, mostrò ben due *loci*, entrambi dell'enzima malico, in grado di discriminare con una accuratezza di oltre il 99 per cento le popolazioni sudoccidentali da quelle alpine. Un terzo *locus*, relativo a un'esterasi, risultò essere anch'esso in grado di discriminare tra i due gruppi di camosci a livello statisticamente significativo. La distanza genetica globale (D = 0,108), calcolata convenzionalmente con la formula di Nei (basata sulla stima della percentuale di sostituzioni alleliche per ogni gene avvenute da quando le popolazioni si sono separate), risultò cadere tra quelle caratterizzanti popolazioni di specie diverse. Inoltre la distanza genetica tra i camosci alpini e quelli pirenaico-appenninici escludeva un flusso genico significativo tra i due gruppi durante le glaciazioni würmiane.

Una distanza genetica molto minore (D = 0,009) venne trovata invece tra il camoscio dei Pirenei e quello appenninico, indicando così che il flusso genico fra queste due popolazioni doveva essere rimasto in atto fino a tempi relativamente recenti (almeno 60 000-40 000 anni fa). Questo era un fatto sorprendente perché l'informazione paleontologica attestava invece la presenza di camosci nell'Olocene delle Alpi Apuane e dell'Appennino settentrionale; era ipotizzabile quindi che lo scambio genetico fra i camosci delle Alpi e quelli dell'Appennino fosse continuato fino a tempi abbastanza prossimi.

Una spiegazione, improbabile ma non impossibile, di quanto sopra era che un nucleo di camosci dei Pirenei fosse stato introdotto in Abruzzo, per esempio nel XVII o XVIII secolo, con finalità venatorie. Questa soluzione (che avrebbe egregiamente spiegato tanto le nettissime affinità tra i camosci pirenaici e quelli d'Abruzzo, quanto la diversità tra questi e i camosci alpini) poteva essere scartata o accettata soltanto sulla base di un'appropriata informazione paleontologica sulla passata distribuzione geografica delle due forme di camosci. Iniziò allora una fruttuosa collaborazione interdisciplinare con Federico Masini, dottorando di ricerca in paleontologia all'Università di Firenze, che già si era occupato dei camosci fossili delle Apuane per la sua tesi di laurea.

Il camoscio appenninico (*Rupicapra pyrenaica ornata*) è l'unico mammifero interamente italiano incluso nella «Lista rossa dei *taxa* animali in pericolo di estinzione» dall'International Union for the Conservation of Nature. Esso è strettamente affine ai camosci dei Monti Cantabrici e dei Pirenei, ma presenta differenze comportamentali, biometriche, elettroforetiche e fenotipiche rispetto al camoscio nordorientale, o alpino, *Rupicapra rupicapra*.

Nel 1968 il noto paleontologo finlandese Björn Kurtén aveva scritto che l'origine del camoscio era «un mistero». Infatti il genere *Rupicapra* compare improvvisamente nel Pleistocene medio (circa 500 000 anni fa), in Francia, associato a elementi faunistici di ambiente freddo come la renna *Rangifer* e l'estinto *Praeovibos*. Tuttavia mancano forme fossili precedenti in grado di aiutarci a ricostruire la sua evoluzione. Il camoscio è un bovide della sottofamiglia *Caprinae*, a sua volta di solito suddivisa in quattro tribù: rupicaprini, ovibovini, budorcatini, caprini. La tribù dei rupicaprini, quella a cui appartiene il camoscio, comprende oggi quattro generi: *Nemorhaedus* e *Capricornis* in Asia meridionale e orientale, *Oreamnos* in Nord America occidentale e *Rupicapra* in Europa centro-meridionale e Asia Minore.

Nel Plio-Pleistocene i rupicaprini erano rappresentati da un ben più ampio numero di generi, ma già nel Pleistocene medio sopravviveva in Europa il solo camoscio (*si veda l'illustrazione a pagina 50 in alto*). I membri di questa tribù non sono mai stati trovati, neanche allo stato fossile, in zone aride, diversificandosi pertanto da ogni altro membro delle *Ca-*

Il repertorio comportamentale del camoscio nordorientale, *Rupicapra rupicapra*, e quello del camoscio sudoccidentale, *Rupicapra pyrenaica*, differiscono per almeno sei moduli etologici. In *a* sono illustrate le posture usate dai maschi delle due specie durante il periodo degli amori, in novembre, per impedire che le femmine si allontanino dall'harem. Nel camoscio sudoccidentale due tipi di comportamento maschile, strettamente collegati alla copula, sono le carezze effettuate con una delle zampe anteriori sul fianco della femmina (*b*) e lo sfregamento del muso contro l'area mammaria (*c*). Quest'ultima postura è la perfetta imitazione del comportamento di un piccolo durante la poppata e apparentemente serve al maschio adulto come una sorta di travestimento per avvicinarsi alla femmina in estro senza suscitarne la ritrosia. L'imitazione è tanto completa da includere perfino i due-tre colpi di muso alle mammelle, vibrati inizialmente dai piccoli per stimolare la secrezione lattea. Un altro modulo etologico tipico del camoscio sudoccidentale è la postura assunta dalle femmine (*d*) durante interazioni minacciose, dove si fondono elementi aggressivi (arcuamento del dorso, erezione pilifera sul garrese, rigidità di movimenti) e di timore (proiezione del muso in avanti). Il camoscio nordorientale, infine, può talvolta corteggiare le femmine toccandole bruscamente con una delle zampe anteriori, tenuta ben tesa. Forse sono state proprio queste differenze etologiche ad avere agito da barriere «pre-copula» agli accoppiamenti misti allorché le due specie di camosci sono venute in contatto durante il Würm.

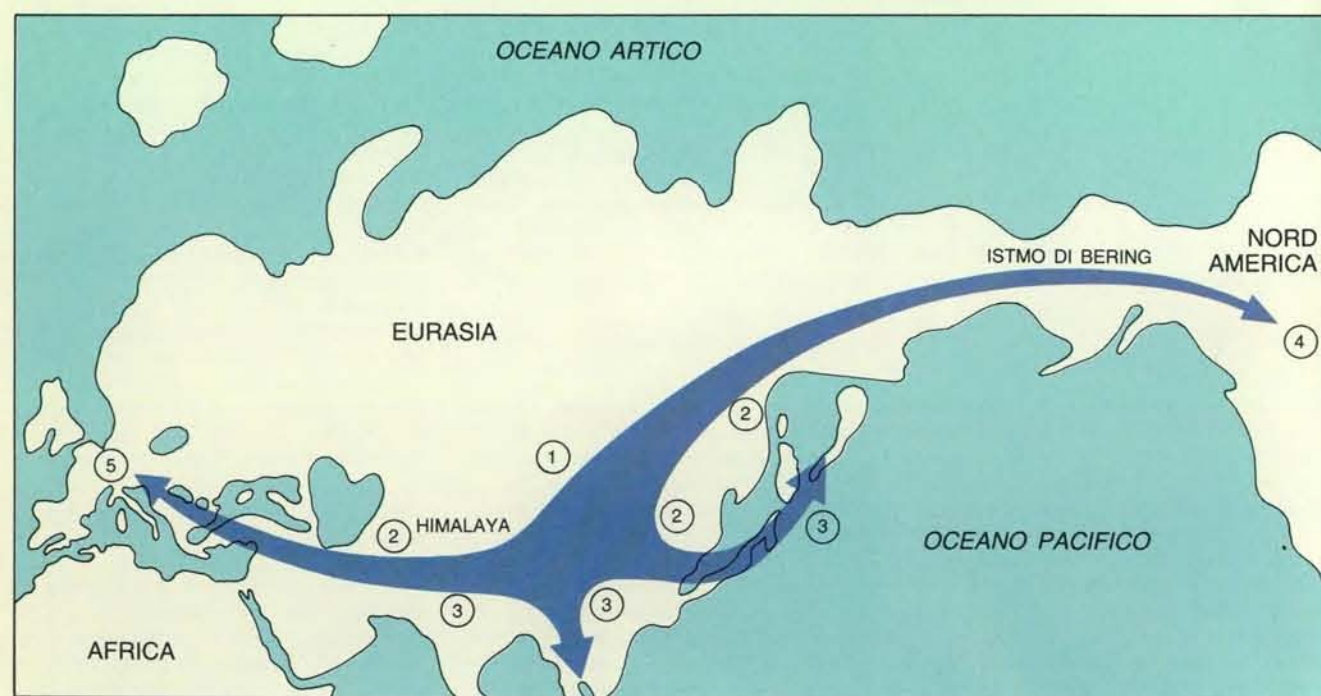
Sono illustrate le caratteristiche dell'ornamentazione del mantello invernale (*a*) e di quello estivo (*b*) del camoscio alpino, *R. rupicapra rupicapra*, e del camoscio appenninico, *R. pyrenaica ornata* (*c*, *d*). Tutte le sottospecie nordorientali mostrano una forte somiglianza con quella alpina, mentre le sottospecie sudoccidentali sono simili a quella appenninica. Le differenze tra sottospecie vertono soprattutto su dimensioni e profilo delle corna e dimensioni corporee.

prinae. Essi sembrano essere stati sempre dipendenti da aree rocciose e montane piuttosto umide, solo raramente discendendo alle altitudini più basse. I fossili dei rupicaprini sono molto rari e dispersi, a differenza di quelli degli altri bovidi, perché il terreno ripido e roccioso preferito da questi ungulati, a causa della forte erosione che comporta, ostacola il processo di fossilizzazione. I depositi lacustri o i depositi in cavità di aree pianeggianti sono infatti i giacimenti fossiliferi migliori. Resta dunque arduo, alla luce della scarsa messe di reperti fossili, tentare una identificazione dei rapporti filogenetici tra le poche forme viventi e quelle fossili. Di fatto i rupicaprini potrebbero essere considerati un gruppo polifiletico.

I rupicaprini esistevano già nel Miocene (oltre cinque-sette milioni di anni fa) tanto in Cina quanto nell'America Settentrionale e, agli inizi del Pliocene (circa cinque milioni di anni fa), troviamo forme insulari in Sardegna (*Nesogoral*) e nelle Baleari (*Myotragus*). Agli inizi del Pliocene troviamo anche in Grecia una sorta di rupicaprino: *Gazella bailloudi*. Il *Myotragus*, la cosiddetta capra delle Baleari, è tra questi senza dubbio la forma più particolare. Di dimensioni simili a una capretta tibetana, aveva enormi incisivi mandibolari sporgenti in avanti per una lunghezza quasi pari a quella di tutto il cranio; probabilmente li usava come scalpelli per raschiare muschi e licheni, suo cibo, da rocce o da tronchi d'albero. Apparentemente esso è stato il rupicaprino fossile di maggior successo, essendo sopravvissuto fino all'Olocene!

A cavallo tra Pliocene e Pleistocene troviamo tre altri rupicaprini: la nordafricana *Numidocapra crassicornis*, da taluni considerata però un alcelafino; il grande *Gallogoral meneghinii*, molto simile all'attuale serau (*Capricornis sumatraensis*) dell'Asia sudorientale, ma presente in Europa occidentale; infine l'antilope-camoscio *Procamptoceras brivatense*, di medie dimensioni, sopravvissuto in Europa orientale fino a circa 800 000 anni fa... appena prima, dunque, della comparsa dei fossili di *Rupicapra*, con cui il *Procamptoceras* è senza dubbio strettamente imparentato. Sarebbe però azzardato ipotizzare una diretta discendenza del camoscio dall'antilope-camoscio, perché questo mostra alcuni caratteri della dentatura e degli assi ossei delle corna, oltre ai parietali più corti, che male si integrano con il ruolo di progenitore del camoscio. Comunque si può concludere che *Numidocapra*, *Procamptoceras* e *Rupicapra* appartengono a una linea filetica già separatasi dagli altri rupicaprini da tre a un milione di anni fa.

La comparsa improvvisa in Europa dei primi fossili di camoscio, dopo che ogni altra forma di rupicaprino sembrava essere scomparsa da questa regione, suggerisce un'immigrazione dall'est avvenuta durante una fase di clima freddo, come dimostra il tipo di fauna a esso associata. È probabile che il camoscio, o il suo diretto avo, sia arrivato in Europa come tardo immigrante nel Pleistocene inferiore/medio, con una delle onde

Dalla distribuzione cronologica dei rupicaprini nella regione europea e nell'Africa settentrionale si vede come i fossili di camoscio compaiano durante le glaciazioni del Riss (Pleistocene medio). Essi sono preceduti da forme di rupicaprini di cui *Numidocapra* e soprattutto *Procamptoceras* sono le uniche relativamente affini al camoscio attuale con il quale formano un gruppo filetico a sé stante.

Nella cartina è approssimativamente indicato l'irradiamento dei vari generi della tribù dei rupicaprini dal Miocene fino ai tempi attuali. Il camoscio ha avuto probabilmente origine e si è evoluto in un'area montuosa situata a ovest della catena dell'Himalaya.

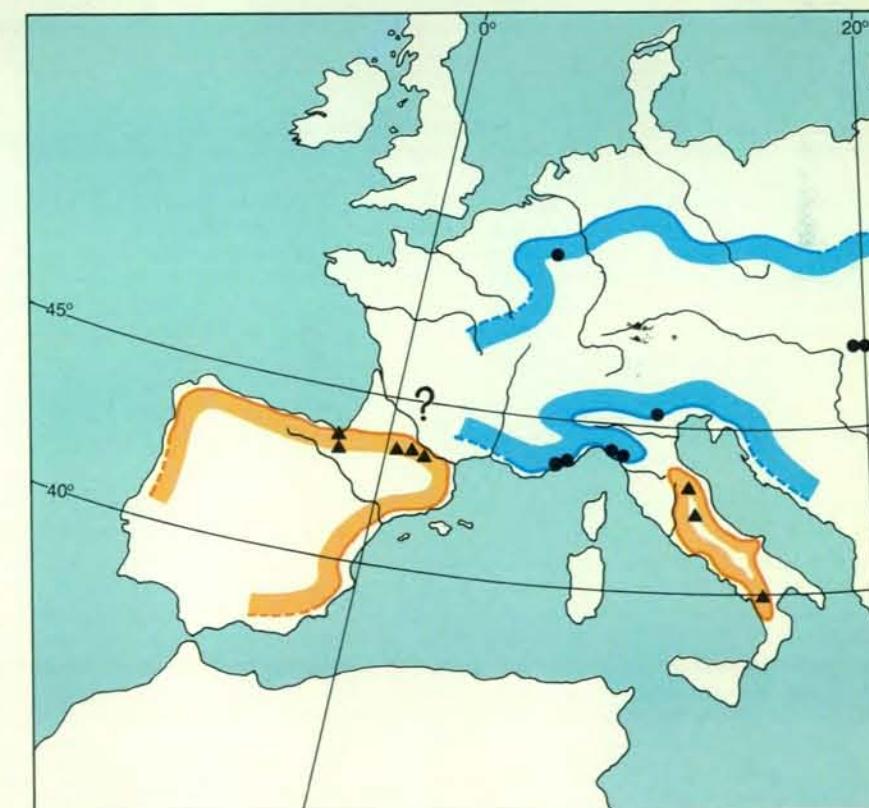
colonizzatrici di fauna tipica di ambienti freddi e aperti che cominciarono a giungere da oriente fin oltre un milione di anni fa. Pertanto si può ipotizzare che una regione montuosa a ovest dell'Himalaya sia stata il centro di dispersione del camoscio o del suo avo diretto, che avrebbe potuto diffondersi verso occidente seguendo le catene montuose del sistema alpino. È meno probabile un'immigrazione attraverso le regioni pianeggianti e steppose dell'Asia occidentale, alla luce della chiara preferenza di *Rupicapra* per le aree rocciose e montane.

Sono soprattutto gli assi ossei delle corna che consentono un'identificazione sistematica delle forme fossili di camoscio, poiché di solito i reperti sono frantumati e ogni altro carattere distintivo è scarsamente attendibile. Tuttavia Masini è riuscito a classificare con sufficiente accuratezza qualitativa alcuni reperti dai quali emerge che camosci del tipo sudoccidentale erano distribuiti anche durante il Würm nella penisola iberica e in quella italiana, mentre altri di tipo alpino avevano una diffusione eminentemente centro-europea, giungendo fino alle Alpi Apuane e agli Appennini settentrionali. Ma non è tutto: un reperto olocenico dell'Appennino umbro, un asse osseo, è talmente simile a quelli attuali dei camosci pirenaici da confermare in pieno l'ipotesi che camosci iberici e appenninici abbiano avuto strettissimi rapporti di parentela. Contemporaneamente l'informazione paleontologica consente di negare un trasporto recente di camosci dai Pirenei all'Appennino, che alla luce dei soli dati relativi ai sistemi gene-enzima sarebbe stato viceversa lecito ipotizzare!

Pertanto, in virtù delle numerose differenze etologiche, biometriche, morfologiche, elettroforetiche, confortate dai dati paleontologici, siamo stati in grado di proporre la separazione dei camosci sudoccidentali in una specie a sé: *Rupicapra pyrenaica*, con due-tre sottospecie, diversa dai camosci nordorientali che invece restano ascritti alla specie *Rupicapra rupicapra* (*si veda l'illustrazione alle pagine 52-53*).

Quale può essere stata la storia evolutiva recente del camoscio? L'ipotesi, che avevo formulato fin dal 1980 in una pubblicazione con Claudio Scala, e che ho recentemente riproposto, con qualche modifica, insieme a Masini suggerisce che il camoscio più antico, quello presente in Europa già durante le glaciazioni del Riss (250 000-150 000 anni fa), appartenesse al tipo sudoccidentale oppure a un suo avo diretto, mentre i camosci di tipo alpino si sarebbero differenziati in qualche fredda area montana dell'Europa nordorientale o in Asia Minore durante un periodo interglaciale o interstadiale.

Non più tardi dell'inizio del Würm (80 000-60 000 anni fa) questi camosci ben adattati al clima freddo, sospinti dal-

Il tipo di dispersione e l'abbondanza stagionale delle risorse alimentari influenzano in misura determinante la socialità dei camosci appenninici. Ridotti a poche centinaia, essi mostrano oggi una scarsissima variabilità genetica: è quindi probabile che un eventuale fattore letale possa abbassarne il numero sotto il livello critico per la sopravvivenza della sottospecie.

È indicata la distribuzione geografica nel Quaternario delle due specie di camoscio esistenti nell'Europa occidentale, *Rupicapra pyrenaica* (*in giallo*) e *Rupicapra rupicapra* (*in blu*). I triangolini e i cerchietti identificano i siti di reperimento dei resti fossili di cui è stato possibile determinare la specie d'appartenenza mediante lo studio degli assi ossei delle corna.

La cartina mostra la distribuzione naturale e attuale delle specie e delle sottospecie di camoscio. Il camoscio alpino, *Rupicapra rupicapra rupicapra*, è stato introdotto per scopi venatori anche in alcune aree dove preesistevano altre sottospecie, inquinandone l'identità genetica. Dieci-quindici camosci alpini vennero anche portati agli inizi di questo secolo in Nuova Zelanda, dove oggi se ne contano non meno di 30 000.

l'aumentata aridità delle regioni del Mediterraneo orientale e del Mar Nero, avrebbero colonizzato i Balcani e l'Europa centrale invadendo aree da cui i precedenti camosci di tipo sudoccidentale si erano ritirati. In particolare, la continuità dell'areale di questi ultimi sarebbe stata interrotta a livello degli Appennini settentrionali e delle Alpi occidentali dall'onda di camosci di tipo alpino provenienti da nord-est.

Durante l'Olocene infine i camosci sudoccidentali, ormai isolati nelle due popolazioni iberica e appenninica, non riuscirono più a riespandersi rispettivamente verso est e nord-ovest (forse a causa della presenza dei camosci di tipo alpino e anche dell'accresciuta pressione antropica nelle aree intermedie). Sembra quindi lecito ipotizzare che, prima o dopo, le due forme di camosci siano venute a diretto contatto, ma senza ibridarsi, come la loro attuale distanza genetica attesta. Il numero cromosomico è lo stesso ($2n = 58$) nelle popolazioni pirenaica, appenninica e alpina: è dunque possibile che gli ibridi siano fecondi. Ma in apparenza non ci sono state ibridazioni o, alternativamente, non hanno avuto successo in termini evolutivi. Si è dunque ipotizzato un isolamento riproduttivo «pre-copula», probabilmente derivato da barriere etologiche. Infatti tutte e cinque le differenze comportamentali identificate nel repertorio maschile dei camosci sono connesse al comportamento riproduttivo, inteso come raggruppamento dell'harem e corteggiamento (*si veda* a, b e c *nell'illustrazione di pagina 48*). In particolare, dallo studio eseguito sul camoscio appenninico è emerso che ben due tipi di comportamento sono significativamente associati alla copula, precedendola.

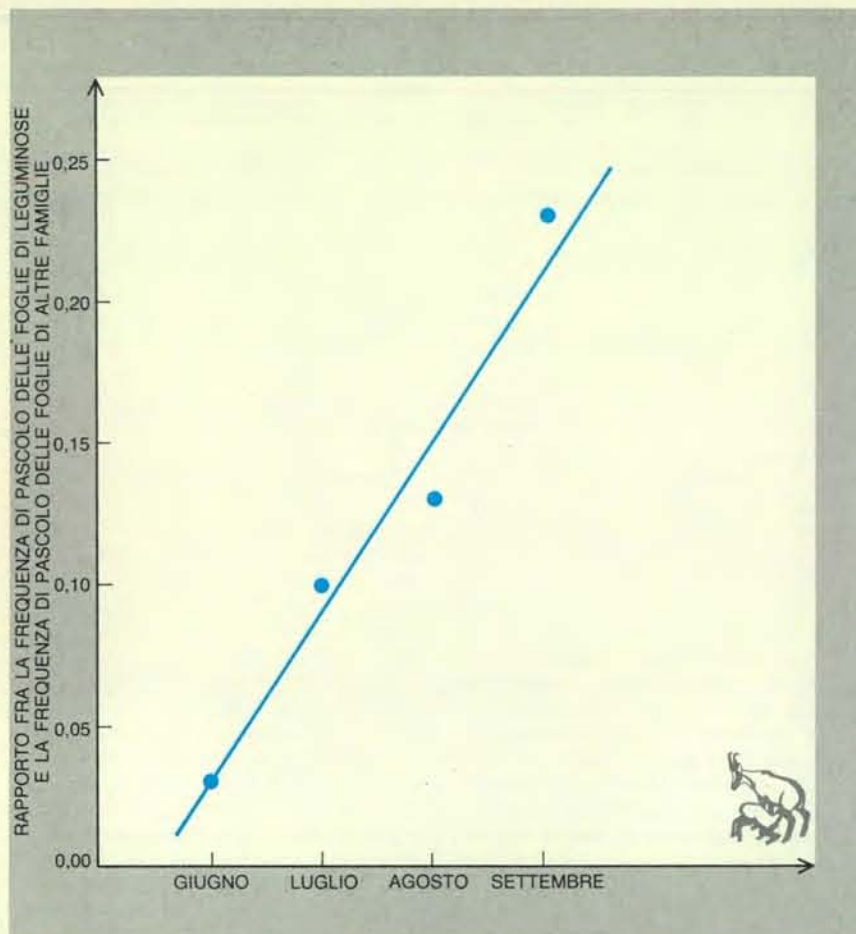
Il fatto che l'unico reperto risalente al Riss, tassonomicamente identificabile e trovato sui Pirenei, sia ascrivibile a *Rupicapra pyrenaica* conforta l'ipotesi della maggiore antichità dei camosci sudoccidentali. Questi inoltre, a differenza di *Rupicapra rupicapra*, hanno in comune con il *Procamptoceras* almeno quattro caratteri degli assi ossei delle corna, a favore della tesi di una loro maggiore arcaicità. Inoltre *Rupicapra rupicapra* mostra due caratteristiche tipiche delle forme adattate ai climi freddi: la struttura anatomica più tozza, con appendici relativamente più corte, che minimizza la dispersione del calore corporeo, e il colore nero del mantello invernale che favorisce l'assorbimento dei raggi solari. Pertanto, pur con qualche approssimazione, tutto sembra quadrare. Sarebbe senza dubbio molto interessante verificare in natura la validità dei meccanismi etologici pre-copula come prevenzione dell'accoppiamento, ma un simile esperimento (che comporterebbe il trasporto di riproduttori di *R. pyrenaica* e di *R. rupicapra* in un biotopo montano pienamente controllabile e ben lontano dagli areali di ambedue le specie per evitare eventuali fenomeni di inquinamento genetico nelle popolazioni naturali) sarebbe molto costoso e difficilmente concretizzabile. Non avrebbe invece alcun senso tentare l'esperimento in cattività, dove le barriere etologiche troverebbero molte più difficoltà ad affermarsi.

Ogni branca della biologia si basa più o meno estesamente sulla sistematica, che ha l'importante funzione di individuare i rapporti filogenetici tra gli organismi. Ascrivere alla stessa specie entità faunistiche così diverse come i due tipi di camosci, quando anche fossero potenzialmente compatibili sul piano riproduttivo, oscurerebbe invece di chiarire i rapporti filogenetici, andando contro proprio alle finalità della sistematica. Del resto non mancano esempi di ibridi fecondi prodotti in natura tra specie diverse di cervidi, canidi e altri mammiferi, che normalmente continuano a restare entità specifiche distinte.

Oggi esistono non meno di 491 500 esemplari selvatici - cioè non in cattività - di *Rupicapra rupicapra* e più di 31 350 individui di *Rupicapra pyrenaica*. Una revisione tassonomica ancora in corso ha assimilato la sottospecie dei Monti Cantabrici, *R. pyrenaica parva*, a quella dei Pirenei, *R. pyrenaica pyrenaica*, e la sottospecie del Caucaso, *R. rupicapra caucasica*, a quella della Turchia, *R. rupicapra asiatica*. Nessuna sottospecie di camoscio si è estinta in tempi recenti, ma quella del massiccio francese della Chartreuse, *R. rupicapra cartusiana*, è fortemente minacciata: ne restano appena un centinaio di individui.

Anche l'ultima popolazione di camoscio appenninico, arroccata su alcune montagne del Parco nazionale d'Abruzzo, corre grossi rischi: ne sopravvivono infatti non più di 350-400 esemplari. Inoltre, mentre il decremento numerico del camoscio della Chartreuse è recente, e quindi la popolazione superstite mostra ancora un'ottima variabilità genetica (con un valore di eterozigosi pari a 0,042), i camosci appenninici devono essere rimasti a densità molto basse per decine, forse centinaia, di anni e soltanto di recente la migliore protezione accordata li ha fatti aumentare di numero. La loro variabilità genetica però è oggi nulla, sia a causa del recente crollo numerico durante la seconda guerra mondiale, quando la scarsa sorveglianza dei guardiaparco determinò un conseguente aumento del bracconaggio, sia soprattutto per la passata permanenza a bassa densità (indirettamente attestata dagli scarsissimi riferimenti a questi camosci nelle cronache di zoologi, cacciatori e viaggiatori degli ultimi secoli, che invece menzionano spesso lupi, orsi, caprioli e cervi).

Questa situazione è pericolosissima per la sopravvivenza del camoscio appenninico: è infatti la variabilità genetica che assicura in una popolazione una risposta differenziata a epidemie o a drastiche variazioni ambientali in generale. Una scarsa eterozigosi può significare che i fattori esterni letali per un individuo saranno con molta probabilità nefasti anche per il resto della popolazione. Ormai non si può più ovviare a questo, ma è possibile diminuire il rischio di una totale estinzione reintroducendo la sottospecie in varie aree idonee dell'Appennino in modo da creare alcuni nuclei alternativi alla popolazione superstite del Parco nazionale d'Abruzzo.

Dal 1913 al 1987 sono state registrate ampie fluttuazioni nel numero dei camosci appenninici presenti nel Parco nazionale d'Abruzzo, dove oggi sopravvive l'ultima popolazione di questa rara sottospecie ridotta a non più di 350-400 esemplari. Le stime numeriche fino al 1970 provengono da documenti conservati negli archivi del Parco, che però non forniscono dati continui. Dal 1970 i guardiaparco censiscono regolarmente la popolazione in autunno. Nel diagramma è stato riportato solo il numero dei camosci effettivamente avvistati, ma si può stimare che circa un quarto della popolazione, soprattutto i maschi adulti che vivono in bosco, sfugga ai censimenti.

Lo studio delle abitudini alimentari dei camosci del Parco nazionale d'Abruzzo è di estrema utilità per valutare l'idoneità dell'ambiente al mantenimento e all'espansione della popolazione. È mostrato il rapporto fra la frequenza di pascolo delle foglie di leguminose e quella delle foglie di altre famiglie in vari mesi. Le leguminose, ricche di proteine e sali minerali anche dopo il periodo vegetativo, vengono predilette dai camosci in estate e in autunno, allorché la qualità delle altre piante è deteriorata. Le abitudini alimentari dei camosci appenninici sono biologicamente ben bilanciate con la disponibilità e la qualità delle risorse naturali e non si rileva impoverimento della vegetazione nemmeno nelle aree a densità più alta (circa 20 camosci per 100 ettari). Gli studi sulla dieta dei camosci appenninici sono stati condotti da ricercatori delle Università di Pavia e di Bologna, coordinati da Carlo Ferrari.

Il camoscio appenninico è un'entità faunistica di grande valore zoologico, sia per l'antichità della sua origine sia per la sua rarità attuale: è infatti l'unico mammifero interamente italiano elencato nella «Lista rossa dei *taxa* animali in pericolo di estinzione» a cura della International Union for the Conservation of Nature (IUCN). Pertanto la sua reintroduzione deve essere programmata con cura particolare in modo da massimizzarne le probabilità di successo. Non è ipotizzabile la sottrazione di capi dall'esigua popolazione abruzzese per un loro successivo rilascio in altre aree appenniniche senza averne prima accertato l'idoneità. Poiché sono la disponibilità e il valore delle risorse alimentari che, con aspetti relativi alla sicurezza da potenziali predatori, determinano la qualità di un'area per le specie erbivore (incluso il camoscio), Carlo Ferrari, professore straordinario di botanica, e Graziano Rossi, dottorando in sistematica ed ecologia vegetale, entrambi all'Università di Pavia, con Claudio Cavani professore associato di zootecnica all'Università di Bologna, hanno condotto uno studio nel Parco nazionale d'Abruzzo per individuare le caratteristiche della dieta dei camosci appenninici in un'area campione ad alta densità numerica di questi rupicaprini, con osservazioni comparative in altre zone del loro areale nel parco.

L'attenzione dei ricercatori si è soprattutto rivolta alla dieta estiva dei branchi di femmine con piccoli, essendo questi di vitale importanza per la prosperità e l'espansione della popolazione, ma è stata raccolta anche qualche informazione sulle scelte alimentari dei maschi. Una differenza tra i sessi, emersa dallo studio coordinato da Ferrari, è che i maschi tendono a fare uso soprattutto delle praterie d'altitudine del tipo *Avena versicolor-Koeleria splendens*, mentre i camosci femmina con i piccoli dipendono quasi interamente da un'altra associazione vegetale, *Festuco-Trifolietum thalii*, comune nell'arco alpino, ma rara ed extrazonale negli Appennini settentrionali e centrali. Questa associazione vegetale è caratteristica di aree con suolo debolmente acido e a prolungato innevamento; nell'ambito del Parco nazionale d'Abruzzo simili requisiti ambientali sono relativamente abbondanti soltanto nella zona con la maggiore densità di camosci (oltre 20 individui per 100 ettari) e, in generale, la consistenza dei branchi appare proporzionale alla locale abbondanza di questa associazione vegetale.

L'importanza dell'associazione *Festuco-Trifolietum thalii* per la sopravvivenza del camoscio appenninico sta nella particolare ricchezza proteica che la caratterizza in primavera e agli inizi dell'estate, proprio nel periodo dell'allattamento e svezzamento dei piccoli. Si può dunque ipotizzare che essa consenta un'ottima dieta per le femmine (e indirettamente anche per i piccoli) in una fase così delicata del loro ciclo annuale.

Una scoperta importante per la comprensione di quanto sia perfettamente bilanciato il rapporto tra camosci e *Festuco-Trifolietum thalii* consiste nell'avere notato che nel corso dell'estate e dell'autunno i camosci modificano la propria alimentazione al variare dei contenuti proteici in fiori e foglie dei vegetali brucati in modo da mantenere approssimativamente immutata la ricchezza della dieta. L'intero comportamento alimentare viene così adattato alle contingenti variazioni mensili nella qualità delle risorse alimentari, al fine di trarre da queste il massimo vantaggio. Elisabetta Bruno, una mia collaboratrice, ha dimostrato infatti che all'impoverimento qualitativo della vegetazione disponibile sulle praterie d'altitudine dall'estate all'autunno (nella stessa area di studio di Carlo Ferrari) corrisponde una crescente competizione alimentare tra i membri del branco e una maggiore selezione nei confronti della vegetazione brucata. In effetti via via che decresce la disponibilità di giovani foglie e fiori, principali fonti delle proteine vegetali assorbite agli inizi dell'estate, nella dieta dei camosci aumenta l'uso di foglie di leguminose. Infatti le proteine di queste foglie hanno un'elevata digeribilità ed esercitano un effetto stimolatore della ruminazione anche quando sono in un avanzato stadio vegetativo. Perciò le leguminose assumono un ruolo sempre più importante nella dieta, venendo attivamente ricercate dai camosci quando diminuiscono le risorse alimentari alternative.

In inverno i branchi si disgregano e i camosci, spinti dalla neve che copre i pascoli, scendono da quote di 1700-2000 metri a 1000-1300 metri di altitudine, scegliendo pendici boscose dove la neve non possa accumularsi per la ripidezza del terreno o per la forza del vento e quindi il cibo sia maggiormente reperibile. Anche in questo caso sono dunque la disponibilità e la qualità delle risorse alimentari che plasmano la scelta dell'*habitat* e la socialità del camoscio. Né durante la permanenza estiva nelle praterie d'altitudine, né durante quella invernale nei boschi ripidi, i camosci sono soliti allontanarsi più di 100-200 metri da aree rocciose e dirupate, dove cercano rifugio se inseguiti da predatori.

È dunque palese l'importanza di una scelta oculata delle zone di potenziale (re)introduzione dei camosci appenninici: infatti, da quanto detto finora, emerge che non tutto quello che è verde è necessariamente un *habitat* idoneo per sostenere una popolazione di questi rupicaprini.

Qualche area dalle appropriate caratteristiche ambientali è già stata identificata, però, tanto sulla Majella quanto sul Gran Sasso: il futuro dei camosci appenninici si trova ora soprattutto nelle mani dei politici e degli amministratori locali.

BIBLIOGRAFIA

COUTURIER MARCEL, *Le chamois*, Arthaud, Grenoble, 1938.

LOVARI SANDRO e SCALA CLAUDIO, *Revision of* Rupicapra *Genus. I. A Statistical Re-evaluation of Couturier's Data on the Morphometry of Six Subspecies of Chamois* in «Bollettino di Zoologia», 47, pp. 113-124, 1980.

LOVARI SANDRO, *Behavioural Repertoire of the Abruzzo Chamois* Rupicapra pyrenaica ornata *(Neumann, 1899) (Artiodactyla: Bovidae)* in «Säugetierkundliche Mitteilungen», 32, pp. 113-136, 1985.

NASCETTI GIUSEPPE, LOVARI SANDRO, LANFRANCHI PAOLO, BERDUCOU CLAUDE, MATTIUCCI STEFANIA, ROSSI LUCA e BULLINI LUCIANO, *Revision of* Rupicapra *genus. III. Electrophoretic Studies Demonstrating Species Distinction of Chamois Populations of the Alps from those of the Apennines and Pyrenees* in *The Biology and Management of Mountain Ungulates*, a cura di Sandro Lovari, Croom-Helm, Londra, pp. 56-62, 1985.

FERRARI CARLO, ROSSI GRAZIANO e CAVANI CLAUDIO, *Summer Food Habits and Quality of Female, Kid and Subadult Apennine Chamois*, Rupicapra pyrenaica ornata *(Neumann, 1899)* in «Zeitschrift für Säugetierkunde», 53, pp. 170-177, 1988.

MASINI FEDERICO e LOVARI SANDRO, *Systematics, Phylogenetic Relationships and Dispersal of the Chamois*, Rupicapra *spp.* in «Quaternary Research», 30, pp. 339-349, 1988.

Carboidrati e depressione

Molti disordini del comportamento fra loro collegati, riconosciuti nello scorso decennio, sono caratterizzati da fame ansiosa e da variazioni di umore; uno dei meglio conosciuti è il disturbo affettivo stagionale o SAD

di Richard J. Wurtman e Judith J. Wurtman

Il 16 maggio 1898 Frederick A. Cook, intrepido esploratore delle regioni polari, fece la seguente annotazione sul suo diario: «L'inverno e l'oscurità si sono lentamente ma inesorabilmente stesi sopra di noi... Non è difficile leggere sui volti dei miei compagni pensieri e stati d'animo... La cortina di oscurità che è caduta sopra il desolato mondo di ghiaccio che ci circonda è discesa anche sul mondo interiore delle nostre anime. Intorno ai tavoli... gli uomini siedono tristi e demoralizzati, persi in sogni malinconici da cui talvolta qualcuno emerge in un vano tentativo di entusiasmo. Per brevi momenti qualcuno cerca di rompere l'incantesimo con una battuta, detta forse per la cinquantesima volta. Altri cercano di ostentare una filosofia ottimistica; ma ogni tentativo di infondere vivide speranze fallisce.»

Oggi sappiamo che i membri della spedizione Cook soffrivano dei sintomi classici della depressione invernale, una patologia legata a una malattia psichiatrica descritta di recente e denominata disturbo affettivo stagionale, o SAD (*seasonal affective disorder*). Come è evidente dal brano del diario, il riconoscimento del nesso tra la depressione e l'inizio dell'inverno non è una novità. Tuttavia, in anni recenti si è registrato un crescente interesse per il SAD e per altri due disturbi del comportamento che hanno in comune con esso alcuni sintomi: l'obesità da consumo eccessivo di carboidrati (CCO, *carbohydrate-craving obesity*) e la sindrome premestruale (PMS, *premenstrual syndrome*). I sintomi comprendono depressione, apatia e incapacità di concentrazione, associate a occasionali periodi di ipernutrizione con eccessivo aumento di peso; essi tendono a essere ciclici e ricorrono in determinati momenti della giornata (di solito nel tardo pomeriggio o alla sera nella CCO), del mese (nei giorni che precedono la mestruazione nella PMS) o dell'anno (generalmente in autunno e inverno nel SAD).

Nel corso dell'ultimo decennio sono state raccolte numerose informazioni che chiariscono non solo le manifestazioni cliniche di questo gruppo di disturbi dell'umore e dell'appetito, ma anche le anomalie dei processi biochimici che ne sono la causa. Attualmente si ritiene che questi disturbi siano influenzati dalle alterazioni biochimiche di due distinti sistemi biologici. Uno di essi coinvolge la melatonina, un ormone che influenza la sensazione soggettiva di benessere e di «energia»; l'altro chiama in causa la serotonina, un neurotrasmettitore che regola il bisogno di ingerire cibi ricchi di carboidrati. Entrambi i sistemi sono influenzati dalla fotoperiodicità, il ciclo giornaliero luce-buio sulla Terra. In effetti, la fotoperiodicità sembra essere determinante nell'andamento ciclico di tutti e tre i tipi di disturbo.

Alle alte latitudini negli emisferi settentrionale e meridionale il SAD compare nel tardo autunno o all'inizio dell'inverno e dura fino alla primavera. Una volta manifestatosi, tende a ripresentarsi ogni anno, a meno che il paziente non si trasferisca in un luogo in cui la durata del giorno non varia significativamente con le stagioni. I pazienti lamentano episodici momenti di depressione, associati a fame ansiosa di cibi ricchi di carboidrati. Si coricano presto e rimangono a letto per 9 o 10 ore, al contrario dei pazienti affetti da depressione non stagionale, che hanno difficoltà a dormire. Il sonno di questi soggetti, tuttavia, è intermittente e non del tutto ristoratore; durante il giorno sono spesso assonnati e hanno problemi di concentrazione. Con l'arrivo della primavera le persone affette da SAD diventano piene di energia e creatività e mostrano una gioia di vivere quasi frenetica. Contemporaneamente, il loro bisogno di carboidrati diminuisce e la maggior parte di essi perde il peso che aveva acquistato durante l'inverno.

Il caso che ora esporremo è esemplare. La paziente M., un'insegnante di 53 anni, è alta un metro e 63 centimetri e pesa 82 chilogrammi. Crucciata per i suoi problemi di peso, per diversi anni ha speso soldi a palate in diete di breve durata. «So che il mio problema sono i carboidrati: quando sono a dieta evito pane, patate e dolci e riesco sempre a dimagrire. Ma quando non sono a dieta divento ansiosa e tesa verso la metà del pomeriggio e non riesco a concentrarmi su quello che sto facendo. Voglio mangiare qualcosa per calmarmi, così compro cracker o ciambelline e li rosicchio. A casa talvolta continuo a mangiucchiare fin quando vado a letto.» A partire dal giorno del Ringraziamento (che negli Stati Uniti cade il quarto giovedì di novembre), la paziente M. attraversò un periodo di stanchezza e depressione durato due mesi. «Dicevo a mio marito di lasciarmi sola e assegnavo esercizi ai miei alunni in modo da non dover parlare in classe. La casa era un disastro. Avevo smesso di mangiare, a eccezione di pane e pasta, ma continuavo ad aumentare di peso. Finalmente, quando venne la primavera, mi sentii meglio, forse perché l'anno scolastico stava per finire e l'estate era vicina.»

I sintomi descritti dalla paziente sono virtualmente gli stessi associati alla CCO e alla PMS, tranne per il fatto che la fame ansiosa di carboidrati si manifesta ogni giorno, di solito nel tardo pomeriggio o nella prima serata, mentre le donne che soffrono di PMS ne sono affette mensilmente, durante la fase luteinica del ciclo ovarico.

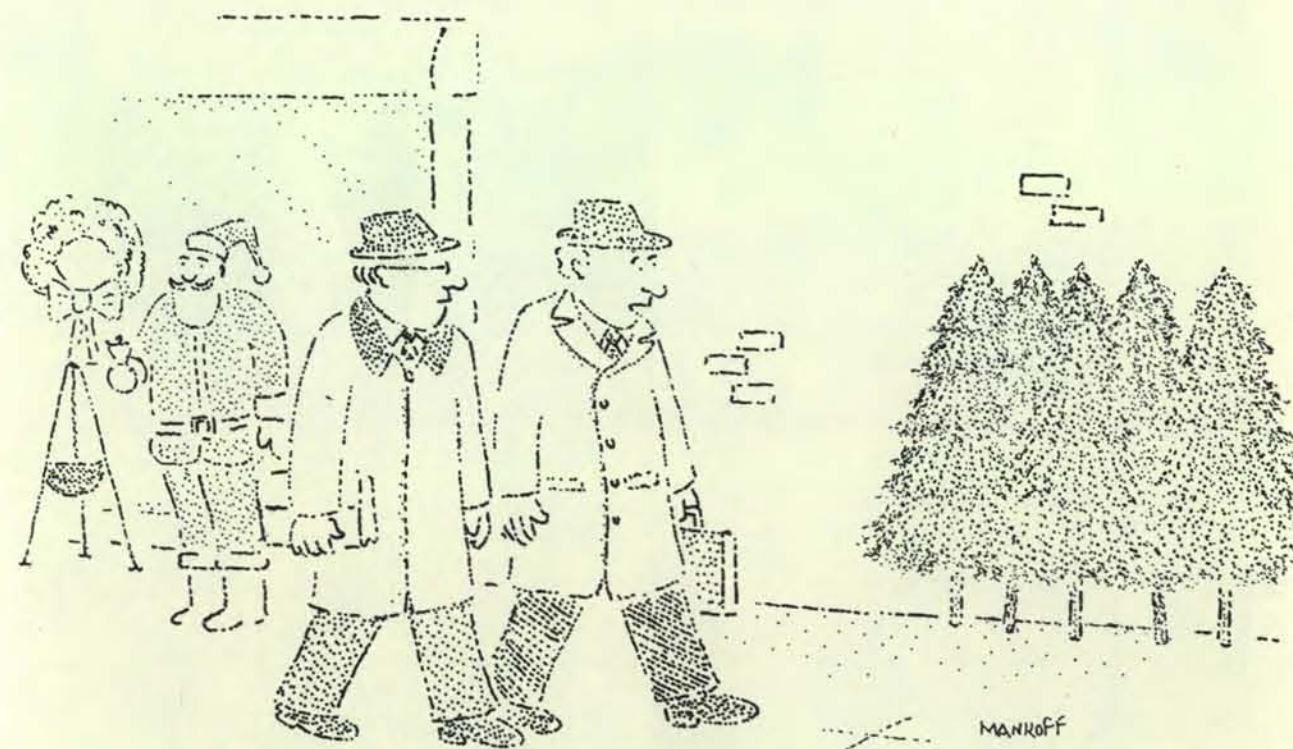
L'interesse per i disturbi stagionali del comportamento si può far risalire agli inizi degli anni ottanta, quando Peter S. Mueller, psichiatra del National Institute of Mental Health, riconsiderò i dati relativi a una donna di 29 anni che aveva avuto come paziente per episodi ciclici di depressione invernale. Nel corso degli anni la paziente si era trasferita in diverse città. Mueller aveva mantenuto i contatti con lei e aveva osservato che, quanto più a nord ella si spostava, tanto più precocemente si manifestava la depressione in autunno e tanto più a lungo essa durava in primavera. In due occasioni, quando la donna aveva compiuto viaggi in Giamaica durante l'inverno, la depressione era scomparsa nel giro di un paio di giorni dall'arrivo.

Mueller formulò l'ipotesi che la luce solare (o la mancanza di essa) contribuisse in qualche modo alla depressione della donna e decise di sperimentare la fototerapia (una forma di terapia che si era precedentemente dimostrata efficace nella cura di bambini itterici e di pazienti affetti da psoriasi). Al mattino, per alcuni giorni di seguito, egli espose la paziente a un «supplemento» di 2500 lux di luce a spettro completo e constatò che in meno di una settimana era guarita dalla depressione.

Le scoperte di Mueller richiamarono l'attenzione di Norman E. Rosenthal, Thomas A. Wehr e Alfred J. Lewy, anch'essi del NIMH, che erano interessati ai vari aspetti della depressione clinica. Essi iniziarono una ricerca su vasta scala della storia naturale della depressione invernale, reclutando un gran numero di volontari per l'osservazione e la cura. I risultati furono rivelatori e interessanti. Essi confermarono l'effetto terapeutico dell'esposizione alla luce nella cura della depressione invernale. Inoltre i dati diedero modo di dimostrare per la prima volta il legame tra depressione invernale e fame ansiosa di carboidrati.

Uno studio successivo di Steven G. Potkin, Daniel F. Kripke, William Bunney e colleghi dell'Università della California a Irvine fornì dati più completi sulla correlazione fra SAD e latitudine negli Stati Uniti. Un questionario pubblicato sul quotidiano «USA Today» nel marzo 1985 aveva fornito una descrizione del SAD, ma senza fare alcun riferimento alla presunta associazione con la durata del giorno. Ai lettori fu chiesto di rispondere affermativamente o negativamente a 15 frasi che descrivevano aspetti ritenuti caratteristici della malattia. A coloro che avessero risposto in modo affermativo a otto o più punti (e perciò erano probabilmente affetti da SAD) fu chiesto di inviare il questionario agli autori: 723 persone lo fecero. L'incidenza del SAD in ciascuno Stato fu determinata dividendo il numero di risposte per la media giornaliera di vendite del quotidiano in quello Stato. I risultati indicarono che 100 persone su 100 000 nelle regioni settentrionali degli Stati Uniti erano affette da SAD; nel sud l'incidenza era di meno di sei su 100 000. È tuttavia probabile che queste cifre siano approssimate per difetto poiché le persone affette da SAD hanno generalmente meno voglia di leggere i giornali e di rispondere ai questionari di quanta ne abbiano le persone sane.

Quasi contemporaneamente iniziammo a effettuare ricerche sui disturbi della nutrizione presso il Clinical Research Center (CRC) del Massachusetts Institute of Technology. Uno studio condotto al CRC durava in media due settimane e si concentrava sul consumo di carboidrati in 20 pazienti appartenenti a due gruppi di peso: soggetti moderatamente obesi (dal 20 al 39 per cento al di sopra del peso ideale) e obesi (dal 40 all'80 per cento al di sopra del peso ideale).

Le abitudini alimentari dei soggetti del nostro studio furono controllate da vicino, sia ai pasti regolari sia tra un pasto e l'altro. Il consumo di spuntini fu misurato tramite un distributore automatico collegato a un calcolatore (progettato da J. Trevor Silverstone del St. Bartholomew's Hospital Medical College di Londra) che era in funzione a qualsiasi ora e conteneva una varietà di spuntini che andavano dai biscotti, ricchi di carboidrati, alle sardine, ricche di proteine. Tutte le scelte possibili contenevano circa la stessa quantità di grassi (sei grammi, per esempio) e di calorie (circa 110). I cibi potevano essere prelevati soltanto componendo un particolare numero di accesso su una tastiera collegata

«Sì, mi sento alquanto depresso, ma, tenendo conto della stagione, posso dire di essere abbastanza contento».

L'attenzione verso il disturbo affettivo stagionale, o SAD, è aumentata nel pubblico in anni recenti. Questo tipo di disordine emotivo è ben rappresentato nella vignetta di Robert Mankoff, pubblicata nel numero del 10 dicembre 1984 della rivista «New Yorker».

al calcolatore, che registrava continuamente il numero e il tipo di spuntini scelti da ciascun individuo. Ai partecipanti allo studio fu chiesto di mangiare come avrebbero fatto normalmente e di non farsi problemi per le calorie assunte. La maggior parte di essi collaborò, ritenendo che i dati ottenuti sarebbero risultati utili per risolvere i loro problemi di peso.

Il consumo di cibo durante i pasti regolari fu misurato fornendo senza limitazioni ai partecipanti porzioni di cibo in contenitori pesati ed etichettati, differenziati da un codice di colori e disposti su un tavolo nella sala da pranzo. I vari cibi, così come gli spuntini del distributore, erano diversi per contenuto di proteine e carboidrati, ma si equivalevano nel contenuto di grassi e calorie. Al termine di ciascun pasto un dietologo pesava quanto era avanzato per stabilire per ogni tipo di cibo la quantità consumata da ogni soggetto.

I nostri studi al CRC ci hanno permesso di vagliare e sfatare un certo numero di miti riguardanti l'obesità, specialmente per ciò che concerne la fame ansiosa di carboidrati. Il più diffuso fra questi miti, forse, è l'idea che tutte le persone obese consumino in eccesso ogni cibo gradevole non appena questo sia disponibile. Invece si scopre che coloro che sentono il bisogno di ingerire carboidrati consumano in eccesso solo questi tipi di alimenti e lo fanno solo in particolari momenti della giornata. Ai pasti essi si comportano normalmente, consumando un totale di circa 1940 calorie al giorno. (Una donna adulta consuma dalle 1500 alle 2000 calorie al giorno, un uomo dalle 2200 alle 2700 calorie.) Verso il tardo pomeriggio o nella prima serata, tuttavia, i volontari iniziavano a fare spuntini, consumando fino a 800 o più calorie addizionali per persona al giorno. Un comportamento simile è stato osservato in donne affette da sindrome premestruale: esse aumentavano il loro apporto calorico assumendo, sotto forma di spuntini, circa 460 calorie in più al giorno rispetto a donne non affette da PMS allo stesso stadio del ciclo mestruale.

Fummo anche colpiti nel notare (con l'aiuto del distributore computerizzato) che i pazienti quasi invariabilmente sottovalutavano il consumo di spuntini. Sembra che uno spuntino, se mangiato velocemente, venga facilmente dimenticato, come se «non contasse». Tuttavia, per chi ha problemi di peso, gli spuntini contano. In alcuni casi essi costituiscono il 30 per cento o più dell'apporto calorico di un individuo.

Inoltre scoprimmo che la maggior parte degli spuntini consumati da pazienti affetti da CCO e da PMS era costituita da carboidrati. Nel nostro studio condotto al Clinical Research Center osservammo infatti che più della metà delle persone obese affette da fame ansiosa di carboidrati non sceglieva mai uno spuntino proteico, sebbene la maggioranza mangiasse normalmente cibi ricchi di proteine all'ora dei pasti. Una spiegazione possibile (ma non ancora provata) per questo comportamento nella scelta dei cibi è che nei soggetti affetti da fame ansiosa di carboidrati la capacità di regolare l'apporto nutrizionale sia alterata nel tardo pomeriggio e nella prima serata. In una persona normale il desiderio di qualcosa di dolce non è frequente né ciclico ed è prontamente appagato, per esempio mangiando uno o due biscotti; nel caso di fame ansiosa di carboidrati, invece, il desiderio può restare inappagato fino a che non viene ingerita una decina di biscotti. Ciò fa pensare che vi sia un anomalo funzionamento del meccanismo di retroazione che informa il cervello dell'avvenuta assunzione di carboidrati. Un'altra possibilità è che questi soggetti consumino spuntini non perché siano affamati, ma perché gli alimenti ricchi di carboidrati migliorano il loro stato d'animo.

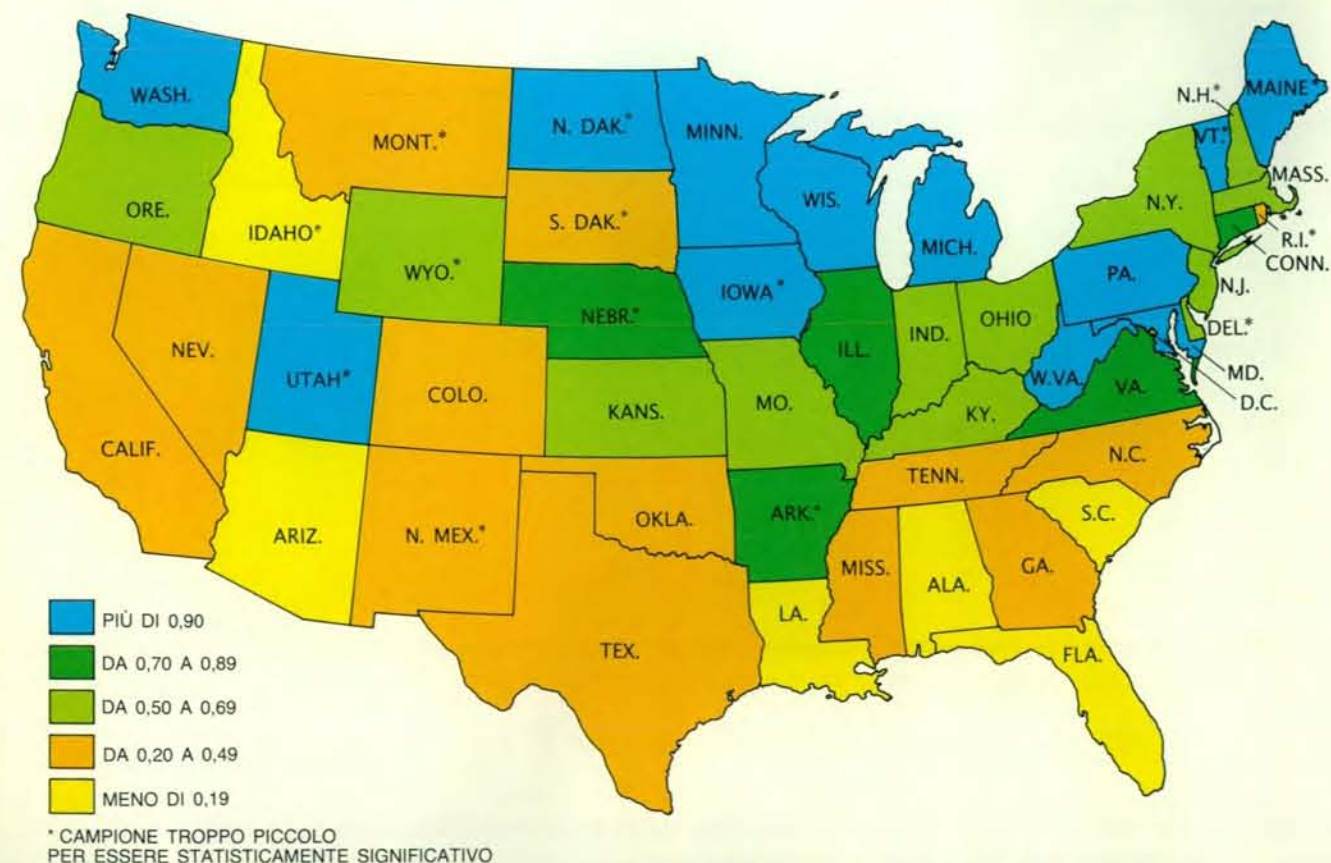
Perché il consumo di spuntini da parte di pazienti affetti da CCO si verifichi in certi momenti della giornata non è chiaro; la ricorrenza ciclica, che è mensile nella PMS e stagionale nel SAD, può riflettere l'azione sul cervello di ormoni ovarici o della melatonina, ma una simile relazione non è ancora stata stabilita per la CCO. È chiaro, in ogni caso, che gli spuntini a base di carboidrati tendono ad accentuare l'obesità perché sono spesso ricchi di grassi e quindi di calorie.

Sembra che la fame ansiosa di carboidrati sia un disturbo dai molteplici aspetti. Fino a due terzi di tutti gli individui obesi sono affetti da fame ansiosa di carboidrati, ma non tutti coloro che presentano questo disturbo sono obesi; molti tengono sotto controllo il proprio peso con l'esercizio fisico, consumando pasti a basso contenuto calorico o soddisfacendo il loro desiderio con carboidrati a basso tenore di grassi, come il popcorn, o con dolci come le gelatine di frutta. Al contrario, non tutte le forme di obesità sono legate a un acuto desiderio di carboidrati. Alcuni soggetti obesi non mostrano alcuna preferenza per i carboidrati, e taluni mangiano eccessivamente soprattutto all'ora dei pasti, ma consumano raramente spuntini.

La nostra ricerca si concentrò anche sulle fluttuazioni di umore fra i soggetti affetti da fame ansiosa di carboidrati. Quando queste persone furono sottoposte a test psichiatrici standardizzati basati o su un'intervista (la scala di Hamilton) o su un questionario a risposta scritta (il Beck Depression Inventory) si manifestò una elevata tendenza alla depressione clinica. Quando a questi soggetti fu chiesto perché essi cedessero a cibi ricchi di carboidrati pur sapendo di aggravare in tal modo la propria obesità, la risposta data fu assai simile a quella fornita dai pazienti affetti da SAD. L'impulso non aveva quasi mai a che fare con l'appetito o con il sapore del cibo; al contrario, quasi tutti affermavano di mangiare per combattere la tensione, l'ansia o l'affaticamento mentale. Dopo aver mangiato, i più riferivano di sentirsi calmi e con la mente libera. Ci chiedemmo quindi se il consumo di quantità eccessive di spuntini a base di carboidrati, che provoca una grave obesità, non potesse costituire una forma di dipendenza da una sostanza; in effetti, il consumo di carboidrati per i loro effetti calmanti e antidepressivi è portato alle estreme conseguenze in questi soggetti, con gravi danni per la salute e l'aspetto dell'individuo.

Con l'aiuto di Harris R. Lieberman e Beverly R. Chew del MIT, uno di noi (Judith Wurtman) iniziò a valutare la relazione tra il consumo di carboidrati e lo stato d'animo. Quarantasei volontari, che comprendevano sia soggetti con fame ansiosa di carboidrati sia soggetti di controllo, furono sottoposti a test psicologici standard prima e dopo aver consumato un pasto ricco di carboidrati e privo di proteine. I primi erano significativamente meno depressi dopo il pasto, mentre gli altri si sentivano affaticati e assonnati. Queste osservazioni inducono a pensare che le persone affette da CCO ingeriscano spuntini a elevato contenuto di carboidrati per ripristinare la loro vitalità affievolita, proprio come alcune persone ricorrono a una tazza di caffè in più quando sentono che il livello di energia o la soglia di attenzione si sta abbassando.

La scoperta che la fame ansiosa di carboidrati, come il SAD, si manifesta con una periodicità precisa, ci ha portati a ritenere che il periodo giornaliero di luce sia in qualche modo legato all'insorgenza ciclica della sensazione di appetito e alle perturbazioni dell'umore. Sapevamo da ricerche svolte circa 25 anni fa che la secrezione di melatonina segue un preciso ritmo circadiano legato ai cambiamenti di luce giornalieri e stagionali, che sembra corrispondere, almeno teoricamente, al ritmo prevalentemente associato al SAD.

La melatonina fu scoperta nel 1958 da Aaron B. Lerner e colleghi della School of Medicine della Yale University, che la isolarono da epifisi di bovino e scoprirono che aveva un effetto schiarente su frammenti di cute di girino. Cinque anni dopo Julius Axelrod e uno di noi (Richard Wurtman), allora al NIMH, ipotizzarono che la melatonina avesse fun-

L'incidenza del SAD negli Stati Uniti varia con la latitudine. In uno Stato settentrionale come il Minnesota il disturbo colpisce più di 100 persone su 100 000, mentre in Florida ne interessa meno di sei su 100 000. Gli asterischi indicano gli Stati dove il campione era troppo piccolo per essere significativo. I dati sono stati raccolti da Steven G. Potkin e colleghi dell'Università della California a Irvine.

Il distributore di spuntini installato presso il Clinical Research Center del Massachusetts Institute of Technology ha fornito dati sulle preferenze alimentari di persone affette da fame ansiosa di carboidrati e di persone sane. Esso contiene alimenti con uguali quantità di grassi e calorie, ma ricchi o di carboidrati o di proteine. Per prelevare uno spuntino la persona deve comporre un numero di accesso alla macchina, che è collegata a un calcolatore. Per ogni persona sono registrati il tipo di spuntino e l'ora in cui esso viene consumato.

zioni ormonali nei mammiferi, per averne osservato la capacità di sopprimere la funzionalità delle gonadi quando veniva iniettata nei ratti. In seguito scoprimmo che la sintesi della melatonina diminuiva quando i ratti erano esposti alla luce e che questo effetto era mediato da interazioni fra la retina, il cervello e particolari nervi simpatici che innervano l'epifisi.

Quasi contemporaneamente, Wilbur B. Quay dell'Università della California a Berkeley dimostrò che i livelli di melatonina nell'epifisi dei ratti presentano un andamento ciclico giornaliero: aumentano di notte e diminuiscono di giorno. Pochi anni dopo, Russell Pelham e colleghi dell'Università di Pittsburgh descrissero fluttuazioni simili dei livelli di melatonina nel plasma umano. Subito dopo, uno di noi (Richard Wurtman) e Harry J. Lynch del MIT scoprirono che i livelli di melatonina nell'urina umana presentano ampie fluttuazioni nel tempo in campioni raccolti dagli stessi soggetti: essi sono almeno cinque volte più elevati di notte che durante il giorno.

Per provare che la variazione temporale dei livelli di melatonina nell'uomo è influenzata dal ciclo giorno-notte, luce-buio, David C. Jimerson del NIMH, Lynch e uno di noi (Richard Wurtman) esaminarono gli effetti di una improvvisa inversione del ciclo luce-buio. Reclutato un certo numero di volontari, ne controllammo le variazioni dei livelli di melatonina nel plasma e nell'urina e ne modificammo quindi il ciclo luce-buio. Li tenemmo in ambiente chiuso e, durante il «giorno» sperimentale, lasciammo le luci accese fino alle 11 del mattino: il loro periodo giornaliero di buio veniva quindi spostato coprendo l'arco di tempo tra le 11 e le 19.

Scoprimmo che i soggetti impiegavano quattro o cinque giorni per adattarsi fisiologicamente al nuovo ciclo di luce, secernendo melatonina quando era buio e sopprimendone la secrezione quando c'era luce. Avevamo dimostrato così che la secrezione di melatonina segue un ritmo circadiano nell'uomo come negli altri mammiferi, che la scansione del ritmo è endogena (generata da un «orologio» cerebrale) e che è regolata dal ciclo luce-buio.

Né noi né altri ricercatori, tuttavia, fummo in grado di dimostrare nell'uomo quello che Axelrod e uno di noi (Richard Wurtman) avevano osservato più di dieci anni prima nei ratti: che la secrezione di melatonina è soppressa quasi completamente se i soggetti sono esposti alla luce durante il periodo di buio del ciclo. Perplessi, giungemmo alla conclusione che l'epifisi dell'uomo è inspiegabilmente insensibile alla luce.

Solo nel 1980 Lewy scoprì che la secrezione di melatonina nell'uomo può essere soppressa quasi totalmente dalla luce, se questa è di intensità sufficiente. Quando i partecipanti al suo esperimento furono svegliati alle due del mattino ed esposti a una luce di 2500 lux di intensità per un'ora e mezza, i loro livelli plasmatici di melatonina diminuirono drasticamente. Perciò la luce ha due effetti sui ritmi della melatonina nell'uomo, così come si verifica nei ratti. Essa può modificare il ritmo della melatonina (come quando il periodo di luce era artificialmente capovolto nel nostro studio sperimentale) o sopprimerne del tutto la secrezione (se viene eliminato il periodo di oscurità). Ciascuna delle due azioni o entrambe potrebbero determinare l'effetto terapeutico della luce nel SAD.

Il lavoro di Mueller, Rosenthal e di altri ha dimostrato che l'esposizione di pazienti affetti da SAD a un'intensa luce addizionale per alcune ore ogni mattina poteva eliminare la depressione e la fame ansiosa di carboidrati dopo pochi giorni. I soggetti obesi affetti da questa forma di ansia non sono ancora stati trattati con la fototerapia, ma uno studio preliminare di Barbara L. Parry del NIMH fa pensare che questa metodologia possa essere efficace nel trattamento di donne affette da PMS, i cui sintomi peggiorano durante l'inverno.

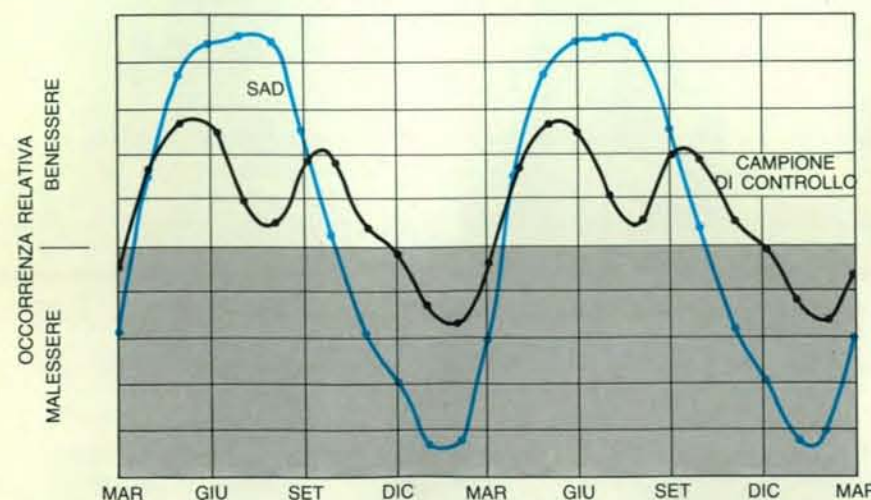
Michael Terman della Columbia University ha scoperto che l'esposizione di persone affette da SAD a una luce di 2500 lux di intensità per due ore ogni mattina porta a una completa remissione sia della depressione, sia della fame ansiosa di carboidrati in circa la metà di essi, di solito dopo pochi giorni di trattamento. La maggior parte degli altri pazienti mostra alcuni miglioramenti, anche se non remissione completa. Sebbene il suo studio non sia ancora ultimato, Terman pensa che sia possibile migliorare l'efficacia del trattamento aumentando l'intensità della luce, per esempio portandola a 10 000 lux, oppure il tempo di esposizione dei pazienti. Certamente questi livelli di illuminazione si avvicinerebbero di più all'intensità della luce del giorno, che varia dai 10 000 lux di una giornata nuvolosa nell'Europa settentrionale agli 80 000 lux di un giorno soleggiato nei pressi dell'equatore. Altri ricercatori, tuttavia, propongono che sia la durata della fototerapia anziché l'orario in cui viene praticata a essere importante nella terapia del SAD. In ogni caso è ora chiaro che la luce deve avere un'intensità di almeno 2500 lux; la normale illuminazione degli interni (che varia per intensità da 250 a 500 lux) non sopprime i sintomi del SAD né inibisce la sintesi di melatonina.

I ricercatori hanno scoperto che la fototerapia è più efficace se viene effettuata al mattino che non durante il resto della giornata. Questa scoperta è stata interpretata da Terman, Lewy e altri come un'indicazione del fatto che la luce fa avanzare il ritmo circadiano dell'individuo e abbrevia la fase di secrezione della melatonina associata all'oscurità. Terman e colleghi hanno notato che la diminuzione dei livelli di melatonina nel plasma, che normalmente si verifica nel primo mattino, è ritardata nelle persone affette da SAD di circa due ore. Forse la luce ad alta intensità provoca una remissione clinica quando l'esposizione avviene al mattino perché in tal modo il periodo giornaliero di secrezione di melatonina viene abbreviato di diverse ore.

Il SAD è causato da un'eccessiva quantità di melatonina o da una sua secrezione troppo prolungata? O la melatonina è soltanto un indicatore di qualche altro processo che sta alla base della malattia? Al momento non siamo in grado di rispondere a queste domande, ma prove circostanziali suggeriscono in effetti un legame diretto tra melatonina e SAD. Lieberman, Lynch e uno di noi (Richard Wurtman) scoprirono che la somministrazione di dosi piuttosto elevate di melatonina a individui sani induce sonnolenza, diminuisce l'attenzione e rallenta i tempi di reazione. Forse l'inizio della secrezione di melatonina alla sera è un importante promotore del sonno, sensibilizzando il cervello ad altri fattori che inducono il sonno. Questo può spiegare perché le persone affette da SAD dormano molto a lungo durante l'inverno, quando il periodo giornaliero di oscurità è quasi doppio di quanto sia in primavera. Un legame tra melatonina e tono dell'umore è suggerito anche dal fatto che la melatonina assunta oralmente aggrava la depressione; purtroppo non è stato sviluppato alcun farmaco in grado di bloccare selettivamente la produzione di melatonina o i suoi effetti.

Ma perché i pazienti affetti da SAD, CCO e PMS hanno tendenza a divorare cibi ad alto contenuto di carboidrati? Perché solo alcune persone sono vulnerabili alla CCO? E come fa il cervello a «sapere» che sono stati o che devono essere consumati carboidrati? Gli abitanti dei paesi sviluppati abitualmente introducono dal 12 al 14 per cento dell'apporto calorico giornaliero sotto forma di proteine e circa tre o quattro volte questa quantità sotto forma di carboidrati. Come viene regolato un fenomeno di questo tipo? Ora sappiamo che la risposta a queste domande coinvolge un neurotrasmettitore, la serotonina. (I neurotrasmettitori sono sostanze liberate da un neurone che, durante la scarica, trasmettono l'impulso nervoso al neurone attiguo attraverso la sinapsi.)

La serotonina è un derivato del triptofano, un amminoacido che normalmente è presente in circolo a bassi livelli. Il tasso di conversione è influenzato dalla quantità di carboidrati nella dieta di un individuo: i carboidrati stimolano la secrezione di insulina, che facilita il passaggio della maggior parte degli amminoacidi nei tessuti periferici, quali i muscoli. I livelli ematici di triptofano, tuttavia, non sono influenzati dall'insulina e quindi la percentuale di triptofano nel sangue rispetto agli altri amminoacidi aumenta quando vengono assunti carboidrati. Dal momento che il triptofano compete con altri amminoacidi per il trasporto attraverso la barriera ematoencefalica, la secrezione di insulina accelera il suo ingresso nel sistema nervoso centrale, dove entra, oltre che in altre cellule, in un particolare aggregato di neuroni, i nuclei del rafe. Qui viene trasformato in serotonina.

Il livello di serotonina, a sua volta, rientra in un meccanismo di retroazione che influenza la quantità di carboidrati che un individuo assume (si veda l'articolo *Nutrienti che modificano la funzione cerebrale* di Richard J. Wurtman in «Le Scienze» n. 166, giugno 1982). Quando il meccanismo di retroazione è alterato, come pensiamo avvenga ciclicamente nei pazienti affetti da SAD, CCO e PMS, il cervello non reagisce all'ingestione di carboidrati e il bisogno di questo alimento continua più a lungo del dovuto.

La serotonina condiziona inoltre il tono dell'umore e lo stato di sonnolenza. Bonnie Spring, attualmente alla University of Health Sciences/Chicago Medical School, ha osservato che le donne che non sono affette da fame ansiosa di carboidrati diventano assonnate e commettono facilmente errori dopo aver consumato un pasto ricco di carboidrati (che ci si aspetta incrementi i livelli cerebrali di serotonina). Risposte simili furono notate in individui obesi non affetti da fame ansiosa di carboidrati da Lieberman e da uno di noi (Judith Wurtman). Al contrario i soggetti che sentivano il bisogno insopprimibile di carboidrati apparivano ristorati e rinvigoriti dopo aver consumato un pasto ricco di queste sostanze.

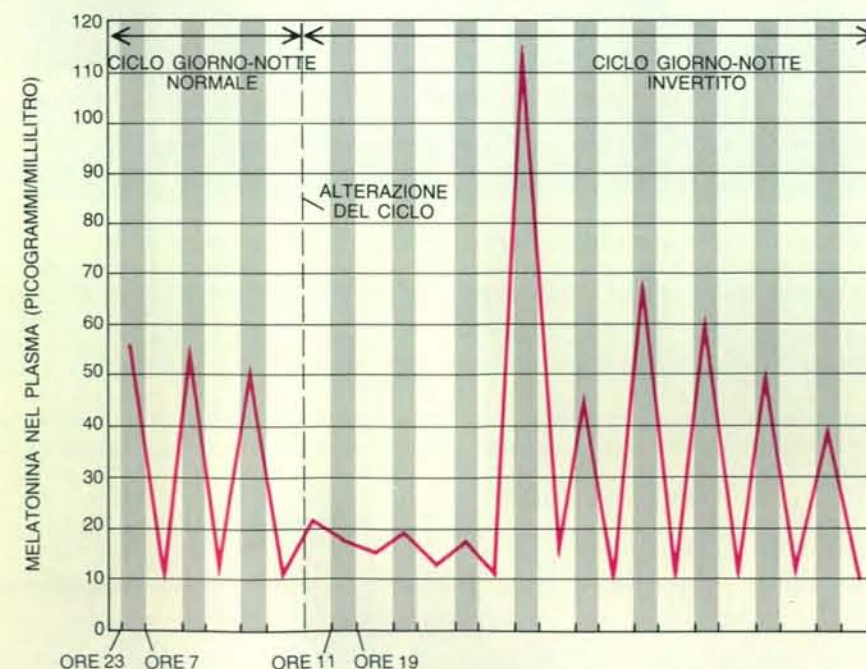
I meccanismi che influenzano le percentuali relative di carboidrati e proteine nella dieta di un individuo sono più evidenti quando i circuiti di retroazione sono soppressi, come accade quando a un paziente vengono somministrati farmaci che influenzano la neurotrasmissione mediata dalla serotonina. I ratti ai quali sia permesso scegliere fra due o più cibi sintetici che contengono proporzioni diverse di carboidrati e proteine normalmente alternano i due alimenti. Se tuttavia ai ratti viene iniettata direttamente serotonina nel cervello o vengono loro somministrati farmaci che aumen-

Le variazioni stagionali di umore sono abbastanza comuni fra gli abitanti di New York (e di altre regioni relativamente settentrionali), ma sono più nette nei pazienti affetti da SAD. I dati sono tratti da uno studio di Michael Terman del New York State Psychiatric Institute.

La fototerapia ha l'effetto di alleviare la depressione e la fame ansiosa di carboidrati associati al SAD. I pazienti che al mattino vengono esposti a luce di elevata intensità per 45-60 minuti mostrano miglioramenti dopo soli due o tre giorni di trattamento.

tano l'effetto della serotonina promuovendo la sua liberazione nelle sinapsi nervose, prolungandone l'attività o stimolandone i recettori, allora l'assunzione di carboidrati da parte dei ratti da esperimento è selettivamente ridotta.

Esperimenti effettuati sull'uomo dimostrano che un farmaco simile alla serotonina, la *d*-fenfluramina (che stimola la liberazione di serotonina nelle sinapsi cerebrali e poi ne prolunga l'effetto bloccandone la riassunzione nel neurone presinaptico) ha un'azione analoga, dato che sopprime selettivamente la fame ansiosa di carboidrati in pazienti affetti da CCO. Abbiamo anche scoperto, in collaborazione con Dermot A. O'Rourke, uno psichiatra del Massachusetts General Hospital, che la *d*-fenfluramina può essere efficace nel trattamento del SAD: essa riduce la fame ansiosa di carboidrati e il conseguente aumento di peso, alleviando contemporaneamente i sintomi della depressione. Più recentemente, in un lavoro condotto insieme ad Amnon Brzezinski della Hebrew University-Hadassah Medical School di Gerusalemme, abbiamo scoperto che la *d*-fenfluramina può essere efficace anche nel trattamento di sintomi analoghi in pazienti affette da PMS. In 12 su 17 soggetti studiati la somministrazione del farmaco per un periodo di sei mesi ha portato alla riduzione sia della fame ansiosa di carboidrati sia della depressione.

Le proporzioni di calorie e di carboidrati consumati sotto forma di spuntini dalle persone affette da SAD variano enormemente da stagione a stagione. In autunno i pazienti assumono più di 3000 calorie al giorno, 1200 delle quali sono date dagli spuntini; in primavera il loro apporto calorico totale scende a meno di 2000 calorie, di cui meno di 500 vengono dagli spuntini (*a sinistra*). Un andamento simile è evidente nel consumo di carboidrati: in autunno quasi il 50 per cento dei carboidrati consumati in una giornata proviene dagli spuntini, mentre in primavera la proporzione si riduce al 30 per cento circa (*a destra*).

La secrezione di melatonina segue un ritmo circadiano sia nell'uomo sia in altri mammiferi. Durante il giorno (*colonne in bianco*) la secrezione dell'ormone dall'epifisi è soppressa e il livello di melatonina nel plasma è basso. Di notte (*colonne in grigio*) la melatonina viene secreta dall'epifisi e il suo livello nel plasma aumenta. Se il ciclo giornaliero di luce viene improvvisamente spostato di 12 ore in modo che il periodo di oscurità vada dalle 11 alle 19 (anziché dalle 23 alle 7) sono necessari parecchi giorni perché il ritmo della secrezione di melatonina si conformi al nuovo ciclo luce-buio. Dopo il periodo di adattamento, si osserva che la secrezione dell'ormone epifisario riprende a seguire un andamento ritmico.

Un'altra malattia che pensiamo possa essere legata alla serotonina (e perciò a SAD, CCO e PMS) è una forma di bulimia associata a un'assunzione smodata di cibi ricchi di carboidrati, ma con vomito scarso o assente rispetto alla bulimia vera e propria. La maggior parte di questi pazienti è costituita da donne moderatamente obese; molte sono gravemente depresse e appartengono a famiglie in cui sono frequenti forme depressive e abuso di alcool. Studi preliminari condotti da G.F.M. Russell dell'Università di Londra e da Arthur G. A. Blouin dell'Università di Ottawa fanno pensare che la *d*-fenfluramina possa essere efficace nella cura di queste donne; quelle che rispondono agli effetti antidepressivi del farmaco hanno maggiori probabilità di beneficiare dei suoi effetti sulla diminuzione dell'appetito.

In generale abbiamo scoperto che i farmaci che agevolano selettivamente la neurotrasmissione mediata dalla serotonina (come *d*-fenfluramina, femossetina, fluossetina, zimelidina e fluvossammina) tendono a provocare una perdita di peso, mentre i farmaci che bloccano la neurotrasmissione mediata dalla serotonina o gli antidepressivi che interagiscono con neurotrasmettitori diversi dalla serotonina hanno l'effetto opposto: spesso inducono fame ansiosa di carboidrati alla quale consegue un aumento di peso.

Nessuno potrebbe sostenere a ragion veduta che i sintomi di SAD, CCO o PMS siano di poco conto. Periodi prolungati di profonda depressione e irritabilità possono compromettere gravemente la capacità di un individuo di intrattenere relazioni umane. Sicuramente, però, non è anormale sentirsi più malinconici durante l'autunno, desiderare cioccolato o pasta, aumentare di qualche chilo ogni inverno o sentirsi irritabili

quando si è assalite dai dolori mestruali.

Infatti i cambiamenti stagionali del comportamento si riscontrano anche in persone sane oltre che in quelle affette da SAD. Su 200 soggetti scelti a caso nell'elenco telefonico della città di New York ed esaminati da Terman e colleghi, la metà affermò di risentire di un calo di energia in autunno e in inverno. Il 47 per cento disse di aumentare di peso in questi mesi, il 31 per cento disse di dormire di più e il 31 per cento di avere un minore interesse per le attività sociali. Fra coloro che dichiararono di subire una perdita di energia in qualche periodo dell'anno, circa il 50 per cento disse che ciò si verificava in autunno e in inverno; solo il 12 per cento dichiarò che ciò avveniva in estate. Terman concluse che una percentuale significativa della popolazione di New York soffre di una forma attenuata di SAD; riteniamo che anche gli abitanti di altre città relativamente settentrionali, come Boston o Minneapolis, possano presentare disturbi analoghi.

A Tromsø, in Norvegia, che trovandosi a una latitudine di 69 gradi non vede il Sole alzarsi sopra l'orizzonte tra il 20 novembre e il 20 gennaio, si ritiene che il 24 per cento della popolazione soffra di insonnia invernale. Charles S. Mullin, Jr., della US Naval Academy ha descritto insonnia diffusa, depressione, irritabilità, alterazione delle facoltà intellettive e un aumento di peso da 9 a 14 chilogrammi circa fra gli scienziati e il personale militare che avevano trascorso l'inverno in Antartide.

Il SAD è dunque semplicemente una forma esasperata della normale risposta dell'uomo al diminuire dei livelli di luce in autunno e in inverno? È forse un analogo del letargo? Probabilmente no. Il letargo negli animali normalmente abbassa la loro temperatura corporea, fa cessare l'attività riproduttiva e fa trascorrere l'inverno in un sonno profondo. Le persone affette da SAD non hanno alcun comportamento simile; anzi, il tempo che trascorrono immersi nel sonno profondo (misurato mediante elettroencefalogramma) diminuisce.

Forse le attuali abitudini di vita aumentano la vulnerabilità alle depressioni stagionali diminuendo la durata del tempo in cui si è esposti alla luce: Daniel Kripke e collaboratori hanno misurato per quanto tempo ogni giorno soggetti anziani in buona salute scelti fra la popolazione di San Diego (una località dal clima particolarmente favorevole) si esponevano al Sole. Sorprendentemente, gli uomini trascorrevano alla luce del Sole soltanto 75 minuti nell'arco di 24 ore, le donne solamente 20 minuti. Non tutti abbiamo bisogno di vivere in California, ma forse la maggior parte di noi necessita di una più prolungata esposizione alla luce, a cui erano abituate le passate generazioni. Forse, così come coloro che lavorano negli uffici si iscrivono a una palestra per ovviare alla mancanza di

La serotonina regola il consumo di carboidrati. Il processo inizia quando un amminoacido, il triptofano (*in arancione*), passa dal circolo sanguigno nel cervello ed entra nei nuclei del rafe. Dopo aver raggiunto un neurone presinaptico, il triptofano è trasformato, tramite un processo in due fasi, in serotonina (*in giallo*). Questa è poi liberata nella fessura sinaptica che separa il neurone presinaptico dal neurone postsinaptico. Quando raggiunge il neurone postsinaptico, la serotonina si lega a particolari recettori. Il livello di serotonina aumenta in risposta al consumo di carboidrati; via via che viene liberata nuova serotonina, ulteriori segnali vengono trasportati al neurone postsinaptico, dove attivano un meccanismo di retroazione. Quando la sua concentrazione è elevata, la serotonina si lega ai recettori presinaptici, impedendo la liberazione di altra serotonina da parte del neurone presinaptico. Essa può anche venire rapidamente rimossa dalla sinapsi per assorbimento nel neurone presinaptico. I farmaci che favoriscono la liberazione di serotonina (*in verde*) o che bloccano il suo riassorbimento (*in blu*) incrementano il trasferimento di segnali attraverso la sinapsi e diminuiscono il consumo di carboidrati; i farmaci che bloccano i recettori postsinaptici della serotonina (*in rosso*) aumentano l'appetito, soprattutto per cibi ricchi di carboidrati.

esercizio, le persone con un'occupazione che si svolge prevalentemente in ambienti chiusi devono garantirsi una adeguata esposizione alla luce.

Vi sono ancora molte cose da chiarire sui disturbi dell'umore e dell'appetito e sul legame tra serotonina e melatonina. Per esempio, perché un paziente affetto da SAD risponde ugualmente bene all'esposizione a una luce addizionale, che presumibilmente agisce influenzando la melatonina, e ai farmaci che stimolano la liberazione di serotonina? E su quale anello della catena di eventi fisiopatologici che conducono al SAD potrebbero agire questi trattamenti? Prima di poter rispondere a queste due domande occorrerebbe sapere se la luce o la melatonina abbiano o meno un effetto diretto sui neuroni che liberano serotonina. Fino a che non avremo risposte più sicure, possiamo almeno rallegrarci per il fatto che questi disturbi possono essere curati con terapie nuove ed efficaci, anche se il meccanismo con cui le terapie agiscono rimane sconosciuto.

BIBLIOGRAFIA

ROSENTHAL NORMAN, SACK DAVID, GILLIN CHRISTIAN, LEWY ALFRED, GOODWIN FREDERICK, DAVENPORT YOLANDE, MUELLER PETER, NEWSOME DAVID e WEHR THOMAS, *Seasonal Affective Disorder: A Description of the Syndrome and Preliminary Findings with Light Therapy* in «Archives of General Psychiatry», 41, n. 1, gennaio 1984.

WURTMAN JUDITH, WURTMAN RICHARD, MARK SHARON, TSAY RITA, GILBERT WILLIAM e GROWDON JOHN, *D-Fenfluramine Selectively Suppresses Carbohydrate Snacking by Obese Subjects* in «The International Journal of Eating Disorders», 4, n. 1, febbraio 1985.

TERMAN MICHAEL, *On the Question of Mechanism in Phototherapy for Seasonal Affective Disorder: Considerations of Clinical Efficacy and Epidemiology* in «Journal of Biological Rhythms», 3, n. 2, 1988.

Alla ricerca del *Proconsul*

Grazie alla recente scoperta di nuovi resti fossili, si è stabilito che questo primate estinto non fu un progenitore dello scimpanzé e del gorilla, bensì l'ultimo antenato comune delle grandi scimmie antropomorfe e dell'uomo

di Alan Walker e Mark Teaford

Il primate estinto *Proconsul* è attualmente il più noto fra i nostri progenitori, ma la storia del suo cammino verso la celebrità a partire dagli eventi piuttosto oscuri connessi con il ritrovamento dei primi resti è ricca di sorprese e di colpi di scena come un romanzo d'appendice. È una storia fatta di episodi bizzarri, nella quale frammenti diversi di esemplari importanti, riportati alla luce, sono separati e spediti in musei sparsi per il mondo, per poi essere riuniti dalla buona sorte decenni più tardi. È anche una storia a lieto fine; spedizioni recenti nei siti degli scavi hanno fornito quasi 800 nuovi esemplari di ominoidei, la superfamiglia dei primati che comprende le grandi scimmie antropomorfe (pongidi), le scimmie antropomorfe minori (ilobatidi o gibboni) e l'uomo. Questi resti hanno molto accresciuto il repertorio di fossili di *Proconsul* e le nuove scoperte dimostrano che questo primate costituisce un valido rappresentante dell'ultimo progenitore comune delle grandi scimmie antropomorfe e dell'uomo.

La storia ebbe inizio nel 1927, quando H. L. Gordon, un colono del Kenya occidentale, trovò alcuni resti fossili mentre stava estraendo calcare da una cava. Pensando che potessero essere importanti, li mandò al paleontologo A. Tindell Hopwood del British Museum. Uno dei fossili non pareva essere altro che un singolo dente sporgente da un nodulo di roccia grande un palmo. Una volta rimossa la matrice, però, l'esemplare risultò essere la mascella sinistra di un ominoideo. Gli altri fossili rinvenuti nella cava permettevano di datare il deposito a circa 18 milioni di anni fa, al Miocene inferiore.

Negli anni venti si conoscevano ben pochi resti fossili di scimmie antropomorfe e nessuno altrettanto antico, cosicché questi esemplari erano estremamente importanti. Hopwood decise tuttavia di non pubblicare la scoperta finché non fosse stato certo di avere veramente identificato un nuovo primate. Dopo essersi procurato fondi per una spedizione, nel 1931 si recò in Kenya, dove riuscì a trovare altri fossili di ominoidei. Due anni dopo Hopwood rese pubbliche le sue scoperte, esprimendo la convinzione che la mascella fossile scoperta da Gordon appartenesse a un nuovo genere, progenitore dello scimpanzé.

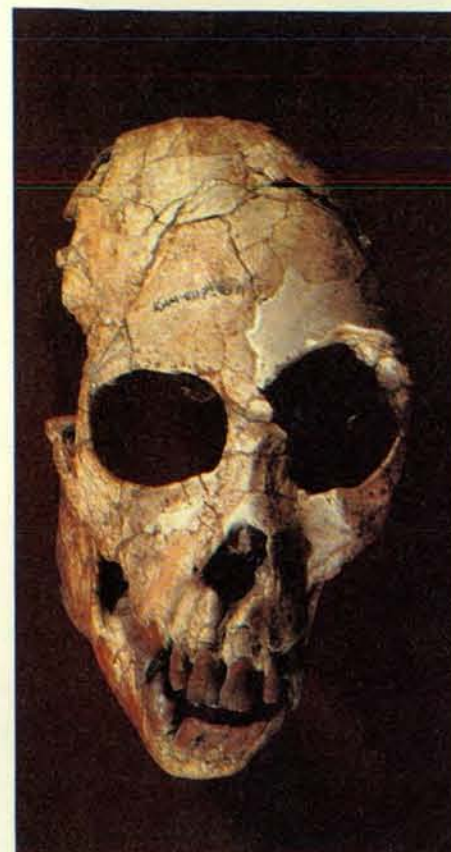
In quegli anni gli appassionati di *vaudeville* a Londra si divertivano a osservare le prodezze di uno scimpanzé che indossava vestiti e cappello, andava in bicicletta e fumava la pipa. Lo scimpanzé si chiamava Consul e, con quella bizzarria di cui spesso danno prova gli scienziati nell'introdurre nuovi nomi nel vocabolario scientifico, Hopwood si ispirò a esso battezzando la nuova scimmia antropomorfa *Proconsul africanus*.

Il capitolo seguente della storia fu scritto da Louis e Mary Leakey, che effettuarono una serie di spedizioni nel Kenya occidentale negli anni quaranta e nei primi anni cinquanta. Sull'isola di Rusinga, nel lago Vittoria, Mary Leakey scoprì nel 1948 un cranio che sarebbe diventato il più famoso fossile di *Proconsul*. Al momento della scoperta, effettuata su un pendio di roccia sedimentaria tenera, la porzione occipitale del cranio era esposta agli elementi ed era erosa in più parti. La faccia e la mascella erano quasi complete, ma la parte posteriore e i lati del cranio erano rappresentati solo da frammenti recuperati ai piedi del pendio.

Si suppose che questo cranio e la mascella trovata da Gordon e descritta da Hopwood appartenessero alla stessa specie. Altre scoperte dei Leakey facevano pensare però che su Rusinga e sulla vicina isola di Mfangano fossero esistite due specie di *Proconsul*: una più grande, delle dimensioni di uno scimpanzé, chiamata *Proconsul nyanzae*, e una più piccola, *Proconsul africanus*, rappresentata dalla mascella di Hopwood e dal cranio rinvenuto nel 1948.

Il successivo esemplare importante fu portato in luce nel 1951 dal geologo Tom Whitworth, durante una prospezione nella zona di Kiakanga, sull'isola di Rusinga. In un corpo verticale di roccia verdastra, di forma grosso modo cilindrica, che attraversava strati di limo grigiastro privi di fossili, Whitworth trovò uno scheletro di suino e altre ossa. Fra quelle che furono recuperate con un paziente lavoro di scalpello dalla durissima matrice di roccia, c'erano frammenti del cranio di un soggetto tra la giovinezza e la maturità, gran parte di un arto anteriore, mano compresa, e frammenti di un piede: tutte le ossa erano appartenute a un singolo esemplare di *Proconsul*.

All'epoca si pensò che il corpo di roccia verdastra, composta da grossolane ceneri vulcaniche depositate dall'acqua, fosse il riempimento di una grande «marmitta», ossia una cavità cilindrica scavata dalle acque di un fiume nella quale i resti degli animali trasportati dalla corrente si sarebbero fossilizzati. In questo caso i ricercatori si sarebbero trovati di fronte a due problemi. Innanzitutto, dato che le ossa e gli scheletri potevano essere stati trasportati nella marmitta anche da notevoli distanze, i resti fossili presenti in essa non rappresentavano necessariamente la comunità animale in cui viveva il *Proconsul*. In secondo luogo, dato che una marmitta viene scavata dall'acqua in sedimenti compatti più antichi, il riempimento della cavità poteva essere molto più giovane della roccia circostante, databile a 18 milioni di anni fa.

Le scoperte effettuate nel 1948 e nel 1951 sono state riesaminate recentemente, assumendo un nuovo ruolo nella storia del *Proconsul*. Questa ripresa di interesse ebbe inizio alcuni anni fa, quando Martin Pickford dell'Institut de Paléontologie di Parigi notò un appunto interessante nel quaderno di campagna di Louis Leakey del 1947. L'appunto si riferiva a frammenti di ossa craniche, forse di primati, rinvenuti nello stesso luogo in cui Mary Leakey aveva trovato il cranio l'anno successivo. Pickford si

Il primo cranio di *Proconsul africanus*, scoperto nel 1948 da Mary Leakey sull'isola di Rusinga, nel lago Vittoria, è mostrato in alto di fronte e di profilo. Altri frammenti vennero aggiunti al cranio nel 1981 da Martin Pickford e da uno degli autori (Walker), che li trovarono in una collezione di resti di tartarughe. Le ossa più scure della mano e del piede di *Proconsul* (*rispettivamente in basso a sinistra e a destra*) vennero rinvenute a Rusinga nel 1951 dal geologo Tom Whitworth; quelle biancastre furono identificate 30 anni dopo presso i National Museums of Kenya. La differenza di colore è dovuta alle diverse sostanze usate per la conservazione dei reperti.

rese subito conto che poteva trattarsi delle parti di cranio che erano state portate via per azione degli agenti atmosferici. Egli riuscì a rintracciare i resti, che erano finiti in una collezione di frammenti di tartarughe conservata presso il National Museum di Nairobi, e stabilì che erano effettivamente le parti mancanti della regione occipitale del cranio di *Proconsul*.

Pickford e uno di noi (Walker) riuscirono a unirli alla parte già nota, completando il cranio dal muso fino alla sommità e, nella parte inferiore, fino al grande foro occipitale.

Avendo ora a disposizione un cranio molto più completo speravamo di conseguire un dato importante dal punto di vista evolutivo: il grado di encefalizzazione del *Proconsul*, cioè il rapporto fra il volume del cervello e il peso del corpo. In effetti siamo riusciti a ottenere una buona stima della capacità cranica di un ominoideo del Miocene. Per valutare la capacità cranica normalmente si misura il volume dell'acqua spostata da un calco dell'interno del cranio. Questo procedimento richiede ovviamente un cranio non deformato; purtroppo l'esemplare in nostro possesso era stato in qualche misura schiacciato e piegato.

Nondimeno, Pickford, Dean Falk della State University of New York ad Albany, Richard J. Smith della Washington University e uno di noi (Walker) escogitarono un modo semplice per eseguire questa stima. Essendo il cranio fatto di materiale non elastico, le lunghezze degli archi al suo interno erano rimaste immutate. Inoltre, la forma generale del cervello del *Proconsul* è simile a quella di alcune scimmie catarrine del Vecchio Mondo, i cercopitecidi, delle quali avevamo a disposizione una serie di calchi dell'interno del cranio. La misurazione degli archi cranici sui calchi rivelò l'esistenza di una relazione statistica fra lunghezza degli archi e capacità cranica nei cercopitecidi. Nel cranio del *Proconsul* riuscimmo a misurare la lunghezza dell'arco che va dall'estremità frontale del cranio fino al margine posteriore del grande foro occipitale. Supponendo che per il *Proconsul* valesse la stessa relazione stabilita per i cercopitecidi, concludemmo che la capacità cranica dell'esemplare fossile in nostro possesso dovesse essere compresa fra 154 e 180 centimetri cubi, con il valore più probabile di 167 centimetri cubi.

Il peso corporeo, dal canto suo, può essere stimato in base a varie misurazioni delle ossa degli arti. Compiendo le opportune misurazioni sulle ossa dell'arto dell'esemplare trovato nel 1951 e su altre ossa a nostra disposizione, potemmo concludere che *Proconsul africanus* aveva un grado di encefalizzazione superiore a quello dei cercopitecidi di dimensioni paragonabili. Riteniamo che una encefalizzazione pronunciata sia una caratteristica delle grandi scimmie antropomorfe attuali, anche se non possiamo esserne certi. I pongidi moderni, che comprendono l'orango, lo scimpanzé e il gorilla, sono molto più grossi delle specie estinte e, benché le dimensioni del cervello aumentino in valore assoluto al crescere della mole corporea, in termini relativi diventano più piccole. Se, però, l'encefalizzazione è un carattere tipico delle grandi scimmie antropomorfe, dovette svilupparsi molto presto nella loro storia evolutiva.

Il cranio rinvenuto nel 1948 ha recentemente permesso di individuare un'altra connessione fra primati antichi e moderni. Sir Wilfrid Le Gros Clark, il paleoantropologo che per primo descrisse il cranio, notò che esso aveva un seno frontale, ossia uno spazio aperto nell'osso frontale del cranio, comunicante con la cavità nasale: questo carattere ha un'importanza non trascurabile, poiché i seni frontali si trovano nell'uomo e nelle scimmie antropomorfe africane, ma non nell'orango, nei cercopitecidi o nelle scimmie antropomorfe minori (gli ilobatidi, che comprendono il siamango e il gibbone). Circa tre anni fa Steven C. Ward della Kent State University stava compiendo uno studio sulle dimensioni e la forma dei seni facciali nei primati superiori. Nel corso di tale studio esaminò il cranio del *Proconsul*. Purtroppo la regione dove erano stati osservati i seni era stata riempita con gesso e poteva essere esaminata solo ai raggi X. A causa delle particolari modalità di fossilizzazione nei sedimenti dell'isola di Rusinga, però, questo metodo non permise di differenziare la matrice di roccia che riempiva la cavità del seno dalla parete ossea e dal gesso. Di conseguenza Ward non fu in grado di verificare l'osservazione di Le Gros Clark.

Recentemente uno di noi (Walker) è riuscito a rimuovere il gesso, mettendo in luce un seno frontale particolarmente grande, riempito di matrice, che si estendeva molto verso la parte posteriore del cranio; questa scoperta ha confermato che queste scimmie antropomorfe del Miocene avevano almeno qualche affinità con i pongidi di oggi, in contrapposizione agli ilobatidi e ai cercopitecidi.

Su Rusinga sono stati ritrovati molti resti di *Proconsul*. Nel 1948 Mary Leakey scoprì il primo cranio nel sito R106, mentre tre anni dopo un altro esemplare fu rinvenuto da Tom Whitworth nel sito R114. Il sito di Kaswanga, scoperto durante la prima spedizione degli autori, ha fornito centinaia di fossili di primati, fra cui almeno nove scheletri di *Proconsul*.

Ancora più complessa è la storia dell'esemplare di *Proconsul* ritrovato nel 1951. Al momento della scoperta, compiuta nel sito R114, le ossa furono estratte da blocchi della dura roccia verdastra che riempiva la marmitta e inviate a Le Gros Clark, che le trasmise a John R. Napier. Napier e il collega Peter R. Davis, entrambi del Royal Free Hospital di Londra, se ne servirono per delineare l'evoluzione degli arti e delle mani dei primati e scrissero una monografia classica sull'argomento. Nel 1964 Napier restituì le ossa al Kenya.

Altre ossa del *Proconsul* furono riunite a queste a causa di un evento fortuito. All'inizio degli anni ottanta fu restituita ai National Museums of Kenya una collezione di suini fossili del Miocene che era stata prestata a un paleontologo. La collezione comprendeva uno scheletro fossile recuperato nella marmitta del sito R114, oltre a un blocco di roccia verdastra contenente numerose ossa articolate che si pensava fossero appartenute anch'esse a un suino. Lo specialista, però, riconosciuto che non si trattava di ossa di suino, le aveva separate.

Durante una visita al museo nel 1980, uno di noi (Walker) esaminò le ossa e si rese conto che appartenevano alla parte inferiore della gamba e al piede dello stesso esemplare di *Proconsul* che era stato descritto da Napier e Davis. L'identificazione fu facilitata dal fatto che l'individuo non aveva ancora raggiunto l'età adulta: molte delle epifisi - le estremità di un osso in crescita, separate dalla

È qui mostrato l'albero genealogico dei primati superiori. La superfamiglia degli ominoidei comprende le scimmie antropomorfe minori o ilobatidi (siamango e gibbone), le grandi scimmie antropomorfe (scimpanzé, gorilla e orango) e l'uomo. Il *Proconsul* visse 18 milioni di anni fa, nel Miocene inferiore. Esso condivide alcuni caratteri con gli attuali gorilla e scimpanzé, ma si tratta per lo più di caratteri poco specializzati; ciò indica che il *Proconsul* fu l'ultimo antenato comune delle grandi scimmie antropomorfe e dell'uomo.

diafisi (la porzione dell'osso compresa fra le due epifisi) da cartilagini di accrescimento - non si erano ancora fuse con il corpo dell'osso, esattamente come si osservava nell'esemplare di Napier. I nuovi reperti permisero inoltre per la prima volta di accertare le proporzioni della mano e del piede del *Proconsul*, dotati di caratteristiche tipiche sia delle scimmie antropomorfe sia dei cercopitecidi, e di dimostrare che questo animale era un quadrupede che si muoveva con lentezza.

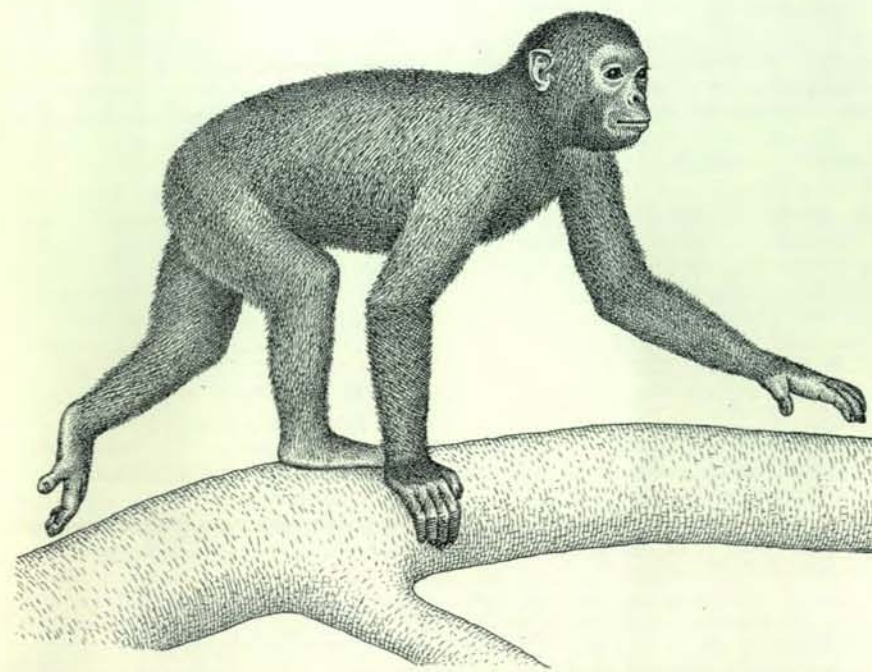
Quanto era accaduto con le ossa di suino indusse Pickford e uno di noi (Walker) a sospettare che altri resti di *Proconsul* potessero essere stati catalogati erroneamente nel museo keniota. In effetti, una volta rintracciate tutte le ossa provenienti dal sito R114, vennero in luce altri resti di *Proconsul*: una scapola, parti di un omero, altre ossa delle mani, le ulne, le tibie e frammenti dei femori. Era ora possibile stimare le proporzioni degli arti dell'animale che rivelarono di nuovo caratteri propri sia delle scimmie antropomorfe (l'arto anteriore), sia dei cercopitecidi (l'arto posteriore).

Se ossa di pongidi e di suini potevano essere mescolate in un classificatore di museo, è possibile che lo fossero anche nella situazione originaria di scavo. Pickford e uno di noi (Walker) cominciarono perciò a cercare ossa di *Proconsul* sfuggite all'attenzione nel sito R114, che dopo il 1951 era rimasto intatto ed era stato ricoperto dalla vegetazione. Essi riuscirono a trovare il sito ma non ebbero il tempo di eseguire sondaggi. Gli scavi da noi compiuti nel sito R114 qualche anno dopo, nel 1984, furono premiati dal ritrovamento della mascella mancante, di mezza clavicola, di un canino, di un importante osso del piede e di un blocco contenente ossa della mano destra, oltre che di svariati fossili di altri mammiferi. Così, dopo oltre trent'anni di separazione, furono finalmente riuniti i diversi frammenti di *Proconsul africanus*, che giacevano in parte nei sedimenti che riempivano la marmitta e in parte nei magazzini del museo dove erano erroneamente catalogati. Il fossile «1951» è oggi il più completo fra quelli degli ominoidei di grandi dimensioni del Miocene.

La spedizione del 1984 fornì importanti informazioni anche sulla struttura della marmitta stessa, informazioni che a loro volta condussero a una migliore comprensione della paleoecologia del *Proconsul*. Varie settimane di scavo nella roccia sedimentaria accanto alla marmitta mostrarono che questa aveva la profondità di almeno quattro metri. I singoli strati di limo non giacevano simmetricamente attorno alla cavità, ma sembrava si fossero depositati in modo non uniforme intorno a un oggetto nella stessa posizione in cui si trova oggi il corpo di roccia verdastra. Le piccole discontinuità nei sedimenti ai due lati della marmitta confermavano questa ipotesi: un oggetto aveva ostacolato il movimento del limo trasportato dalla corrente e infine era caduto, provocando la formazione di discontinuità nei sedimenti. L'ipotesi spiegava anche la deposizione asimmetrica del limo stesso: nella parte a monte dell'ostacolo si erano depositati più ciottoli e ghiaia che nella parte a valle.

Infine divenne chiaro che la presunta marmitta era in realtà qualcosa di completamente diverso. In quella posizione, 18 milioni di anni fa, si trovava un grande albero che era stato parzialmente sepolto da limo e sabbia. L'albero non fu distrutto dal seppellimento, ma probabilmente ne morì e il suo interno divenne cavo. Fu allora occupato da varani, pitoni, pipistrelli e piccoli carnivori, che lasciarono nel tronco le loro ossa e quelle delle loro prede. Uno dei carnivori probabilmente catturò il *Proconsul* e lo trasportò all'interno del tronco cavo per mangiarselo; in effetti alcune delle articolazioni del nostro esemplare mostrano segni di denti. Con il tempo il cavo dell'albero fu riempito da un miscuglio di ossa e sedimenti, che infine si solidificò a formare il corpo cilindrico di roccia conservatosi sino a oggi.

La soluzione del mistero della marmitta permise di chiarire due problemi importanti. Innanzitutto, oggi è certo che la roccia verdastra e i fossili presenti nel deposito sono, in termini geologici,

Sono occorsi 30 anni per ricostruire lo scheletro del *Proconsul*. Le parti trovate al momento della scoperta da parte di Whitworth nel 1951 sono indicate in grigio. Le parti in blu furono identificate nel 1984 da Pickford e Walker nelle collezioni dei National Museums of Kenya; quelle colorate in rosso furono portate in luce dagli autori nel 1984, scavando nel sito R114.

Il *Proconsul* è qui ricostruito come doveva apparire 18 milioni di anni fa. La più piccola delle due specie che vivevano su Rusinga aveva le dimensioni di una femmina di babbuino e arti anteriori e posteriori di lunghezza uguale. I suoi movimenti erano relativamente lenti ed è probabile che non avesse sviluppato specializzazioni per il salto, la brachiazione, la locomozione sulle nocche o la vita al suolo. Gli esemplari della specie più grande avevano dimensioni circa quattro volte superiori a quelli della specie di taglia più piccola.

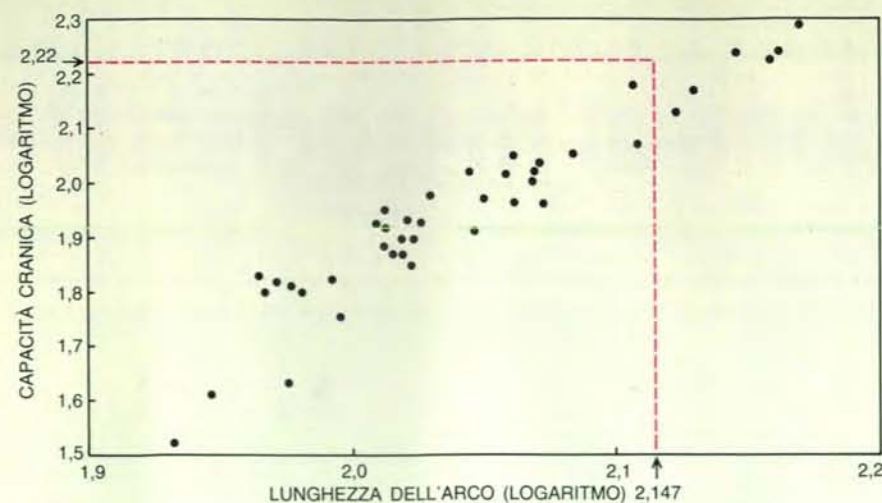
coevi ai limi e alle sabbie circostanti, di colore grigio; questi sedimenti sono stati datati applicando il metodo del potassio-argo alle lave sovrastanti e sottostanti ed è stata loro attribuita un'età di 18 milioni di anni. In secondo luogo, dato che l'albero contiene fossili di animali che lo usarono come posatoio o rifugio, il complesso dei resti rappresenta chiaramente la comunità animale all'interno della quale viveva il *Proconsul*.

I lavori nel sito R114 condussero casualmente a un'altra scoperta importante. Uno dei frequenti temporali tipici del lago Vittoria riempì d'acqua lo scavo fino all'orlo. Non potendo per il momento proseguire i lavori, il personale dei National Museums of Kenya, diretto da Kamoya Kimeu, colse l'occasione per compiere un sopralluogo nelle aree circostanti. Questa ricognizione si rivelò molto fruttuosa: venne scoperto un nuovo sito nel quale, da allora, sono state riportate in luce centinaia di ossa intere e migliaia di frammenti ossei. In effetti il nuovo sito, chiamato oggi Kaswanga Primate Site, si rivelò così spettacolare da indurci a stipulare un accordo con il capo dell'isola di Rusinga per rimandare di cinque anni la destinazione dell'area alla coltura del mais.

Fra gli importanti ritrovamenti compiuti nel sito vi sono almeno nove scheletri interi o parziali di *Proconsul*, che erano stati trasportati dalle acque superficiali in un piccolo fosso formatosi nell'antico terreno miocenico. Essi rappresentano individui di età variabile, da neonati ad adulti; sono probabilmente presenti esemplari di sesso sia maschile sia femminile. Oggi conosciamo praticamente ogni parte dello scheletro del *Proconsul* dai resti di uno o più individui.

Purtroppo le ossa di Kaswanga erano state in gran parte spezzate o scheggiate dal rigonfiamento delle argille e dei limi in cui esse giacevano o dalle radici delle piante. Siamo stati perciò costretti a rimontare i vari esemplari incollandone assieme le parti, un lavoro che non è ancora terminato. Benché l'operazione abbia avuto successo nel caso delle ossa più grandi, è stato difficile distinguere i piccoli frammenti cilindrici delle ossa degli arti di esemplari neonati dalle falangi delle mani e dei piedi di individui giovani e adulti.

I fossili di Kaswanga ci hanno permesso di determinare le proporzioni delle mani, dei piedi e degli arti del *Proconsul*; quando il nostro lavoro di ricostruzione sarà completato, saremo in grado anche di descrivere le varie fasi della crescita del *Proconsul*. Dato che l'analisi dei nuovi reperti richiederà qualche anno, abbiamo deciso di concentrarci innanzitutto sulle ossa già oggetto di precedenti ricerche. Così facendo siamo riusciti a risolvere una controversia concernente l'anatomia funzionale e il significato evolutivo del polso del *Proconsul*.

Il polso delle scimmie antropomorfe e

È qui riportata la capacità cranica dei cercopitecidi attuali in rapporto alla lunghezza dell'arco cranico. La misurazione diretta della capacità cranica del *Proconsul* era impossibile in quanto il cranio ritrovato nel 1948 era schiacciato. La lunghezza degli archi al suo interno tuttavia non era stata modificata dalla deformazione. Inoltre il cervello del *Proconsul* aveva una forma simile a quella dei cercopitecidi attuali, cosicché si può ritenere valida la stessa relazione fra lunghezza degli archi e capacità cranica. Il cranio rinvenuto nel 1948 presenta un arco interno, dalla parte frontale al grande foro occipitale, lungo 140 millimetri. Riportando sull'ascissa il logaritmo di tale valore, si ricava sull'ordinata il logaritmo della capacità cranica dell'esemplare, corrispondente a 167 centimetri cubi.

dell'uomo differisce da quello della maggior parte dei mammiferi, compresi i cercopitecidi: esso non ha una vera articolazione fra l'ulna (una delle ossa dell'avambraccio) e il polso vero e proprio. Nei cercopitecidi invece una piccola estensione ossea dell'ulna poggia in un incavo formato da due ossa del polso, il pisiforme e il piramidale.

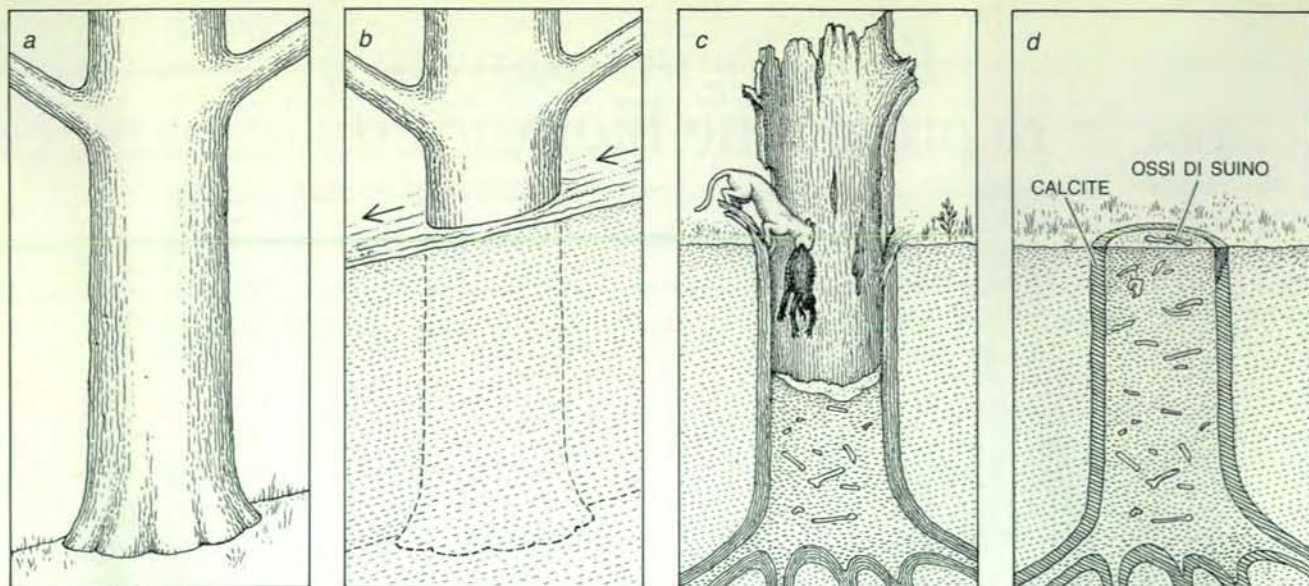
Fra le ossa dello scheletro di *Proconsul* trovate nel 1951 c'erano quelle della mano sinistra, che sono state oggetto di numerose pubblicazioni e tesi di laurea. Il piramidale però era gravemente danneggiato e questo fatto diede origine a divergenze sulla natura dell'articolazione del polso in *Proconsul*. Grazie ai resti trovati a Kaswanga, ora disponiamo di molti nuovi esemplari del pisiforme e del piramidale. In collaborazione con K. Christopher Beard della Johns Hopkins University, siamo riusciti a dimostrare che il polso del *Proconsul* e quello dei cercopitecidi sono simili, dato che in entrambi i casi vi è un'articolazione diretta fra il polso e l'ulna. Benché sotto questo aspetto il polso del *Proconsul* sia primitivo, per altri versi esso preannuncia la maggiore mobilità che si osserva in certe parti del polso degli ominoidei attuali.

Questo mosaico di caratteristiche risulta spesso presente quando si analizzano gli ominoidei del Miocene. Come ha sottolineato Michael D. Rose della University of Medicine and Dentistry of New Jersey a Newark, gli ominoidei del Miocene non assomigliano né ai cercopitecidi né alle grandi scimmie antropomorfe, bensì presentano caratteri propri. La peculiare combinazione di caratteristiche del *Proconsul* rende difficile l'interpretazione funzionale dei fossili: non solo non esistono analoghi moderni dell'anatomia di questo animale, ma ciascuno dei complessi anatomici di *Proconsul* mostra una combinazione di caratteri del tutto particolare. Per esempio, alcune ossa della caviglia sono esili e simili a quelle dei cercopitecidi, ma l'alluce robusto assomiglia a quello dei pongidi. La stessa mescolanza di caratteri si ha nel bacino del *Proconsul*: l'ileo, o porzione superiore, è simile a quello dei cercopitecidi, mentre l'acetabolo (la sede dell'articolazione con la testa del femore) è ampio e poco profondo, come nelle grandi scimmie antropomorfe.

I ricercatori stanno cominciando a rendersi conto, in base a diversi dati, che in passato i progenitori dei cercopitecidi dovettero condurre una vita meno arboricola dei loro discendenti moderni. A sua volta questo ha permesso di capire che molti dei caratteri che il *Proconsul* condivide con i pongidi moderni sono caratteri primitivi che si sono semplicemente conservati per milioni di anni. Questi caratteri non specializzati non indicano strette affinità evolutive. (Dopo tutto, sia l'uomo sia l'opossum hanno cinque dita.) Il bacino del *Proconsul*, per esempio, mostra che questi animali non avevano callosità ischiatiche, i cuscinetti di tessuto indurito sulle natiche su cui siedono i gibboni e i cercopitecidi. La mancanza di callosità ischiatiche, però, è un carattere primitivo; se la presenza di queste strutture è una specializzazione comune a gibboni e cercopitecidi, la loro assenza non è affatto prova di una stretta connessione del *Proconsul* con gli scimpanzé, che ne sono anch'essi privi. Lo scimpanzé ha conservato semplicemente un carattere ancestrale che risale a un'epoca anteriore a quella del *Proconsul*. D'altra parte, il seno frontale è presente nel *Proconsul*, nell'uomo e nei pongidi, ma non nei gibboni e nei cercopitecidi. La condivisione di caratteri specializzati indica che il *Proconsul* è correlato più strettamente ai pongidi moderni che non ai cercopitecidi moderni.

I nostri scavi recenti hanno gettato luce anche sul numero di specie di *Proconsul* che vissero sulle isole di Rusinga e di Mfangano. Molti ricercatori sono d'accordo da tempo sul fatto che la piccola specie di *Proconsul* trovata sulle isole e rappresentata dal cranio scoperto nel 1948 da Mary Leakey non sia da identificare con il *Proconsul africanus* rinvenuto sulla terraferma e rappresentato dalla mascella di Hopwood. Se questo punto di vista è corretto, dovevano esistere almeno tre specie: due di piccole dimensioni e una più grossa, *Proconsul nyanzae*, che è stata trovata anche sulle isole. Altri ricercatori, come Pickford, Jay Kelley della Brown University e David Pilbeam della Harvard University, hanno sostenuto che esistevano solo due specie: gli esemplari di piccole dimensioni rinvenuti sulle isole sarebbero in realtà le femmine di *Proconsul nyanzae*.

Insieme con Christopher B. Ruff, del nostro dipartimento presso la School of Medicine della Johns Hopkins University, siamo stati in grado di verificare l'ipotesi stimando la mole corporea sia degli esemplari grandi sia di quelli piccoli e paragonando le differenze trovate con quelle effettivamente documentate per maschi e femmine di primati viventi. Fra le centinaia di nuovi fossili venuti in luce nel sito di Kaswanga ci sono molti femori di entrambi i tipi di *Proconsul*. Ruff ha dimostrato che è possibile stimare con precisione il peso corporeo di primati quadrupedi misurando la distribuzione del tessuto osseo nella sezione trasversale del femore. Due femori completi di grandi dimensioni e due di piccole dimensioni (che, per quanto possiamo dire in base a tutti gli esemplari di *Proconsul* a noi noti, sono pienamente rappresentativi) forniscono stime del peso corporeo rispettivamente di 37 e 9,6 chilogrammi. In altri termini, il peso delle presunte femmine era pari solo a un quarto del peso dei presunti maschi.

Nessun mammifero terrestre vivente, e tanto meno nessun primate, presenta un dimorfismo sessuale così estremo per quanto riguarda il peso corporeo. Questo fatto ci ha condotto a rifiutare l'idea che le forme grandi e quelle piccole di *Proconsul* venute in luce sull'isola di Rusinga rappresentino maschi e femmine della stessa specie. La nostra conclusione è che, circa 18 milioni di anni fa, su Rusinga e Mfangano vivessero due specie diverse di scimmie antropomorfe pri-

La «marmitta» che conteneva l'esemplare di *Proconsul* scoperto nel 1951 è in realtà, contrariamente alle interpretazioni iniziali, quanto rimane di un tronco d'albero fossilizzato. Nel Miocene inferiore un tronco d'albero posto in riva a un fiume (*a*) fu gradualmente sepolto da sedimenti grigiastri (*b*). A causa del seppellimento l'albero marcì e il suo interno cavo fu usato come tana da vari animali (*c*); il cadavere del *Proconsul* fu probabilmente trasportato nella cavità da un creodonte. Infine il tronco fu colmato da ossi e sedimenti verdastri nei quali i fossili rimasero inclusi fino al 1951 (*d*). Da ultimo la corteccia dell'albero venne sostituita da calcite.

mitive: una specie piccola, diversa da quella di terraferma descritta da Hopwood, e la specie più grande *Proconsul nyanzae*. Sembra che queste specie fossero molto simili per forma e proporzioni, ma diverse per mole corporea.

Le somiglianze di forma e proporzioni indicano importanti analogie di postura e locomozione. Sotto molti aspetti gli esemplari di *Proconsul* che vivevano sulle isole manifestano nelle proporzioni degli arti e nelle superfici articolari una tipologia primitiva; gli arti anteriori e quelli posteriori, per esempio, avevano lunghezza simile. D'altra parte, entrambe le forme erano prive di coda, e questo è un carattere specializzato. Nella loro monografia ormai classica Napier e Davis conclusero, fondandosi su un'analisi delle sole ossa dell'arto anteriore, che il *Proconsul* era un quadrupede agile, saltatore, che si muoveva in modo simile all'attuale entello o langur asiatico e che presentava taluni caratteri scheletrici tipici della brachiazione.

I fossili venuti in luce di recente mostrano, al contrario, che il *Proconsul* era una specie arborea piuttosto lenta, dai movimenti prudenti, che non aveva evidenti specializzazioni per il salto, per la brachiazione, per la locomozione sulle nocche della dita e per una vita terricola.

Lo status del *Proconsul* è mutato considerevolmente nei 60 anni trascorsi da quando Gordon trovò il primo frammento di mascella. Hopwood ritenne che il *Proconsul* fosse un antenato dello scimpanzé, e quest'idea fu ulteriormente sviluppata negli anni sessanta e settanta da alcuni antropologi che videro nelle varie specie di *Proconsul* gli antenati delle diverse specie dei moderni pongidi.

Nell'ultimo decennio, però, è stata scoperta una gran quantità di nuovi materiali, e non solo sull'isola di Rusinga. Richard e Maeve Leakey, dei National Museums of Kenya, per esempio, hanno trovato in un sito del Kenya settentrionale i resti di almeno tre nuovi generi di scimmie antropomorfe risalenti a 17 milioni di anni fa. I nuovi fossili presentano qualche somiglianza con il *Proconsul*, ma anche alcune differenze. La varietà delle scimmie antropomorfe presenti, nel Miocene inferiore, in una regione relativamente piccola dell'Africa orientale era evidentemente molto maggiore di quanto non si ritenesse in passato.

Queste scoperte hanno permesso di dimostrare che le interpretazioni tradizionali dell'evoluzione degli ominoidei primitivi erano semplificazioni grossolane, basate su campioni limitati sia nello spazio sia nel tempo. Questa nozione, insieme al riconoscimento che molte caratteristiche ritenute peculiari degli ominoidei potrebbero essere in realtà comuni a tutti gli antropoidei, ha consentito ai ricercatori di farsi del *Proconsul* un'immagine molto diversa.

Esso non fu un progenitore specializzato dello scimpanzé o del gorilla, anzi ha in realtà ben pochi caratteri specializzati che lo correlino a questi primati moderni. Il *Proconsul* ci appare invece, più genericamente, un progenitore di tutte le grandi scimmie antropomorfe e dell'uomo. Esso è così poco specializzato che, se si confrontasse il *Proconsul* con un teorico progenitore più antico di tutti gli ominoidei, soltanto alcuni caratteri - come il seno frontale e la mancanza di callosità ischiatiche - rivelerebbero che questo primate del Miocene inferiore visse dopo la divergenza fra i pongidi e le altre scimmie antropomorfe.

BIBLIOGRAFIA

NAPIER JOHN R. e DAVIS PETER R., *The Fore-Limb Skeleton and Associated Remains of* Proconsul africanus in *Fossil Mammals of Africa*, 16, pp. 1-69, 1959.

ROSE MICHAEL D., *Miocene Hominoid Postcranial Morphology* in *New Interpretations of Ape and Human Ancestry*, a cura di Russell L. Ciochon e Robert S. Corruccini, Plenum Press, 1983.

WALKER ALAN e PICKFORD MARTIN, *New Postcranial Fossils of* Proconsul africanus *and* Proconsul nyanzae in *New Interpretations of Ape and Human Ancestry*, a cura di Russell L. Ciochon e Robert S. Corruccini, Plenum Press, 1983.

WALKER ALAN, FALK DEAN, SMITH RICHARD e PICKFORD MARTIN, *The Skull of* Proconsul africanus*: Reconstruction and Cranial Capacity* in «Nature», 305, n. 5934, 6 ottobre 1983.

BEARD K. CHRISTOPHER, TEAFORD MARK F. e WALKER ALAN, *New Wrist Bones of* Proconsul africanus *and* P. nyanzae *from Rusinga Island, Kenya* in «Folia Primatologica», 47, n. 2, pp. 97-118, 1986.

Il problema della rete di lunghezza minima

Dato per esempio un insieme di cento punti, qual è la rete di lunghezza minima che li collega? La soluzione di questo affascinante problema ha finora eluso i calcolatori più veloci e le menti matematiche più acute

di Marshall W. Bern e Ronald L. Graham

L'immaginaria Società di servizi telefonici Steiner stimò che avrebbe risparmiato milioni di dollari se avesse potuto trovare il modo di collegare i suoi cento abbonati con la rete telefonica più corta possibile. Per risolvere il problema, la Steiner si rivolse alla Società di calcolo Cavalieri, nota per avere i programmatori e i calcolatori più veloci del mondo. Dopo una settimana la Cavalieri presentò un programma atto a risolvere il problema della Steiner e dimostrò che il programma aveva in effetti trovato, in un'ora soltanto, una rete di lunghezza minima per 15 dei 100 abbonati. La Steiner pagò 1000 dollari per il programma della Cavalieri e s'impegnò a pagare, in ragione di un centesimo al secondo, il tempo che il calcolatore avrebbe impiegato per trovare la soluzione completa. Ma quando il calcolatore ebbe finito il calcolo per tutti i 100 abbonati, la società di servizi telefonici doveva alla Cavalieri migliaia di miliardi di dollari per le spese di calcolo, e nel frattempo i suoi abbonati si erano tutti spostati di molti chilometri... o per decisione autonoma o a causa della deriva dei continenti!

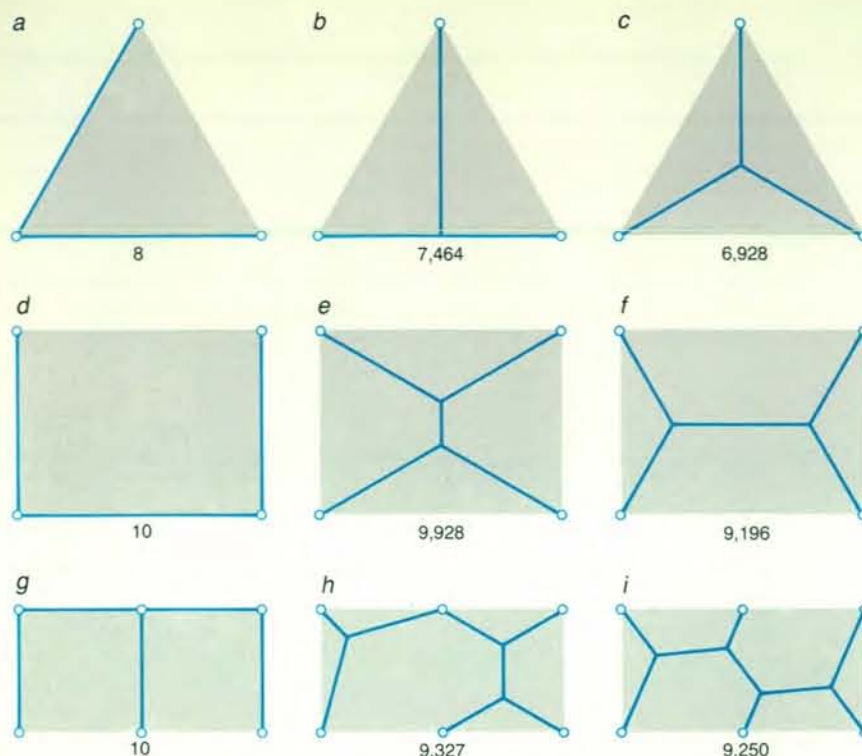
Il programma venduto dalla Cavalieri alla Steiner era scadente? Questo dilemma è un esempio del problema di Steiner, che consiste nel trovare la rete di segmenti di lunghezza complessiva minima che colleghi un insieme di punti assegnato. Il problema di Steiner non può essere risolto tracciando semplicemente segmenti fra i punti assegnati, ma se ne può trovare una soluzione introducendo altri punti, i punti di Steiner, che servono da raccordi nella rete di lunghezza minima. I matematici e i teorici dell'informatica hanno costruito algoritmi, cioè procedure precise, per determinare l'ubicazione e il numero dei punti di Steiner. Ma neppure i migliori di questi algoritmi eseguiti sui calcolatori più rapidi riescono a fornire la soluzione per un insieme consistente di punti assegnati, perché il tempo necessario per risolvere un problema del genere sarebbe troppo lungo. Inoltre il problema di Steiner fa parte di una classe di problemi per i quali oggi molti teorici ritengono che non si potrà mai trovare un algoritmo efficiente. Per questo motivo la Società di calcolo Cavalieri dovrebbe essere scusata.

D'altra parte la Cavalieri avrebbe potuto costruire un programma più conveniente dal punto di vista pratico, capace di fornire soluzioni un poco più lunghe della rete minima. Soluzioni approssimate del problema della rete minima vengono calcolate comunemente per poi essere impiegate in numerose applicazioni, tra cui ricordiamo la progettazione di circuiti integrati, la determinazione dell'albero evolutivo di un gruppo di organismi e la riduzione al minimo dei materiali impiegati per reti di tubature, telefoniche e stradali.

Nella sua formulazione più generale, il problema di Steiner fece la sua prima comparsa nel 1934, in un articolo di Miloš Kössler e Vojtěch Jarník, ma acquisì notorietà solo nel 1941, quando Richard Courant e Herbert E. Robbins lo inserirono nel loro libro *Che cos'è la matematica?*. Courant e Robbins collegarono il problema alle ricerche di Jakob Steiner, un matematico del secolo scorso che lavorò all'Università di Berlino, che aveva studiato il seguente problema: trovare le coordinate di un punto che renda minima la somma delle distanze tra il punto stesso e i punti di un insieme assegnato. Un caso particolare del problema su cui Steiner aveva lavorato era stato peraltro enunciato per la prima volta già verso il 1640: trovare le coordinate di un punto P che renda minima la somma delle distanze fra P e tre punti assegnati. Il problema fu risolto indipendentemente da Evangelista Torricelli e Francesco Cavalieri. Torricelli e Cavalieri dedussero che se gli angoli nel punto P misurano tutti almeno 120 gradi, allora la somma delle distanze viene resa minima.

Sfruttando il fatto che gli angoli in P misurano almeno 120 gradi, Torricelli e Cavalieri fornirono una costruzione geometrica per trovare P (*si veda l'illustrazione a pagina 83 in alto*). Si tracciano i segmenti che collegano i punti dati (chiamati A, B e C, dove B è il vertice dell'angolo più grande). Se l'angolo in B è maggiore o uguale a 120 gradi, allora il punto P coincide con B. In altre parole la rete minima è semplicemente costituita dai segmenti che congiungono A con B e B con C. Se l'angolo in B misura meno di 120 gradi, il punto P giace entro il triangolo. Per trovarlo si traccia un triangolo equilatero sul lato maggiore del triangolo, quello tra A e C, in modo che il terzo vertice X del triangolo equilatero sia opposto al punto B. Si inscrive il triangolo equilatero in una circonferenza e si congiunge B con X. Il punto cercato P è individuato dall'intersezione di questo segmento con la circonferenza. Congiungendo P con i punti A, B e C si ottengono tre angoli di 120 gradi e si ha la rete di lunghezza minima. Inoltre la lunghezza di questa rete coincide con la lunghezza del segmento BX. In questo articolo chiameremo X punto sostituto perché sostituendo i punti A e C con X la lunghezza della rete resta immutata.

Il problema di Steiner con tre punti e quello con molti punti hanno in comune molte proprietà. Le soluzioni hanno una struttura ad albero; questo significa che sopprimendo un segmento qualunque della rete minima, almeno uno dei punti resta isolato. In altre parole non è possibile percorrere la rete partendo da un punto e tornando allo stesso punto senza ripercorrere qualche segmento. Pertanto le soluzioni del problema con tre punti e di quello con molti punti si chiamano alberi di Steiner, i segmenti si chiamano rami e i punti che (come P) debbono

Un calcolatore a bolle di sapone (*in alto*) sfida un calcolatore elettronico (*in basso*) nella ricerca della rete di connessioni di lunghezza minima che collega 29 città. Il calcolatore a bolle di sapone, in cui la disposizione geografica delle città è simulata dall'ubicazione dei cavicchi, rende minima la lunghezza delle pellicole di sapone in una zona locale. Esso fornisce una rete corta, ma non necessariamente la più corta. Il calcolatore elettronico esegue un algoritmo compilato da Ernest J. Cockayne e Denton E. Hewgill dell'Università di Victoria, che fornisce con certezza la rete minima. Il problema con 29 punti è prossimo al limite attuale delle capacità di elaborazione.

Il problema della rete ha diverse soluzioni per punti collocati ai vertici di un triangolo equilatero, di un rettangolo e su una «scala». In *a*, *d* e *g* i punti sono collegati senza l'aggiunta di punti suppletivi e la soluzione ottenuta costituisce un albero ricoprente minimo. In *b*, *c*, *e*, *f*, *h* e *i* sono illustrati alberi di Steiner, che si ottengono aggiungendo punti di raccordo suppletivi. Solo *c*, *f* e *i* sono alberi di Steiner, o reti, di lunghezza minima. Il numero sotto ciascuna soluzione fornisce la lunghezza totale approssimativa dei segmenti.

essere aggiunti per costruire l'albero prendono il nome di punti di Steiner.

Il problema di Steiner con tre punti fornisce informazioni anche sulla geometria degli alberi di Steiner di lunghezza minima. In primo luogo ogni angolo misura almeno 120 gradi, il che implica che ciascuno dei punti assegnati è collegato all'albero da non più di tre rami. In secondo luogo in ogni punto di Steiner si incontrano esattamente tre rami, ad angoli di 120 gradi. In terzo luogo il numero dei rami di un albero è sempre pari al numero dei punti complessivi, assegnati e di Steiner, meno uno. Infine, poiché in ogni punto di Steiner s'incontrano esattamente tre rami e almeno un ramo deve toccare ciascuno dei punti assegnati, in qualunque problema il numero massimo di punti di Steiner è pari al numero dei punti assegnati meno due.

Dati un certo numero e una certa distribuzione di punti assegnati, sono molti gli alberi di Steiner distinti che godono di queste proprietà. Alcuni di questi alberi, denominati soluzioni localmente minime, non possono essere resi più corti con una piccola perturbazione, per esempio uno spostamento esiguo di un ramo o la suddivisione di un punto di Steiner. Ma non tutti gli alberi di Steiner localmente minimi sono soluzioni di lunghezza minima. Può accadere che per trasformare una rete in un albero di lunghezza minima, che si chiama albero di Steiner globalmente minimo, sia necessario operare ridistribuzioni su ampia scala di punti di Steiner.

Possiamo illustrare tutto ciò mediante un insieme di punti assegnati che corrispondono ai quattro vertici di un rettangolo avente lati di tre e quattro metri. Le soluzioni contengono due punti di Steiner, che possono essere disposti in due modi diversi. Ciascuna disposizione dà luogo a un albero di Steiner avente tre rami, che formano angoli di 120 gradi, collegati a ciascun punto di Steiner. Se i punti di Steiner sono disposti nel senso della larghezza, l'albero di Steiner localmente minimo che ne risulta ha una lunghezza di circa 9,9 metri. Se i punti di Steiner sono disposti nel senso della lunghezza, ne risulta un albero di Steiner globalmente minimo che misura circa 9,2 metri.

Per poter trovare una rete minima si può adottare un'impostazione sistematica, consistente nell'esaminare tutti gli alberi di Steiner localmente minimi, calcolare la loro lunghezza e scegliere quello più corto. Tuttavia, poiché i punti di Steiner possono essere collocati ovunque, non è scontato che tutti gli alberi di Steiner localmente minimi si possano calcolare in un tempo finito. Z. A. Melzak dell'Università della British Columbia ha superato questa difficoltà e ha costruito il primo algoritmo per il problema di Steiner.

Nell'algoritmo di Melzak si considerano molti possibili collegamenti tra i punti assegnati e molte possibili collocazioni per i punti di Steiner. L'algoritmo è costituito grosso modo da due parti: nella prima parte si individuano semplicemente tutti i possibili sottoinsiemi dell'insieme dei punti assegnati. Nella seconda parte si generano numerosi alberi di Steiner possibili per ciascun sottoinsieme, sfruttando una costruzione simile a quella impiegata per risolvere il problema con tre punti.

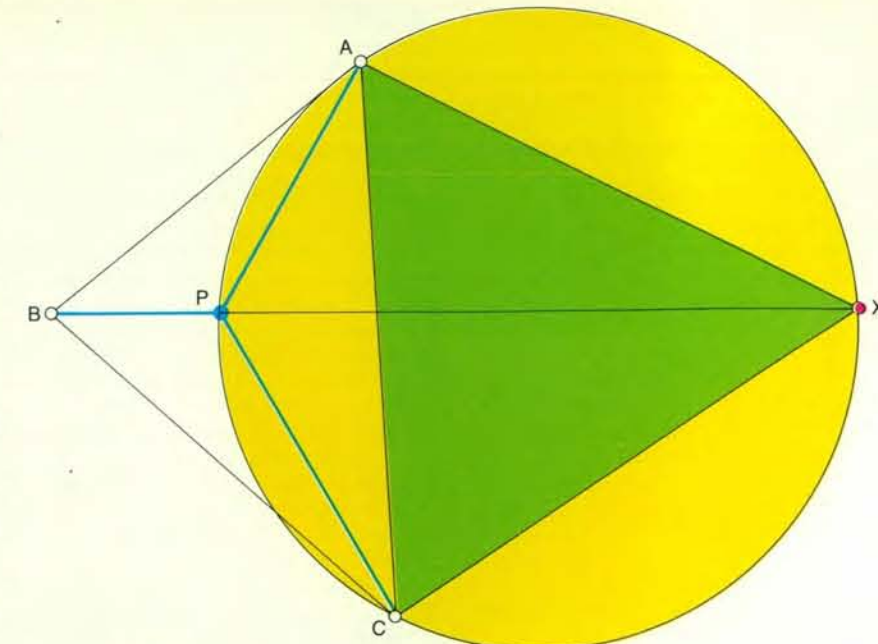
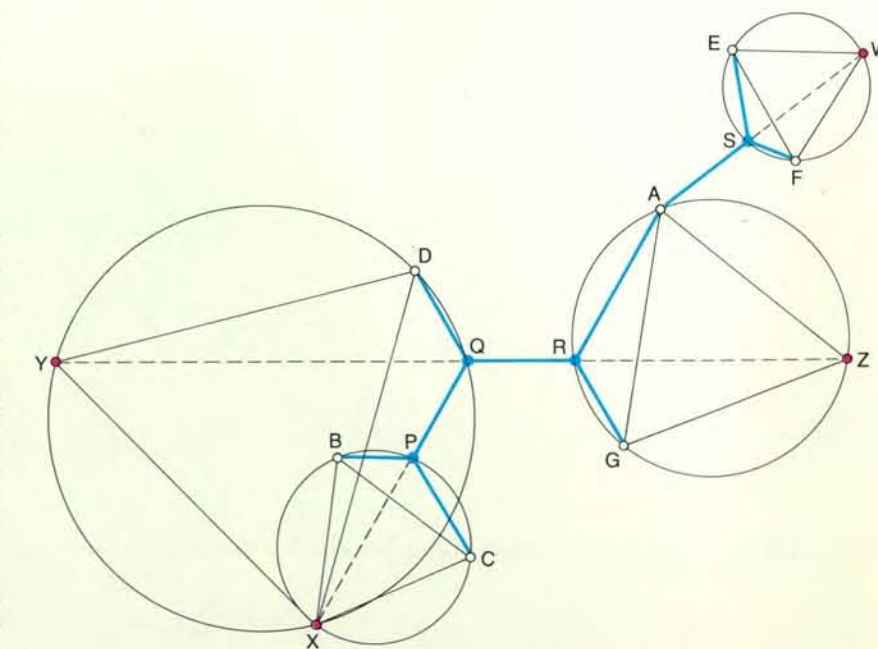
Come nel problema con tre punti, un punto sostituto può prendere il posto di due punti assegnati senza che la lunghezza della soluzione vari. Nel problema generale, tuttavia, l'algoritmo deve scegliere la coppia da sostituire e lo fa provando tutte le scelte possibili. Inoltre il punto sostituto può essere collocato o da una parte o dall'altra del segmento che congiunge la coppia, perché il triangolo equilatero impiegato nella costruzione può essere orientato secondo due direzioni. Dopo che una coppia di punti del sottoinsieme è stata sostituita da uno dei due possibili punti sostituti, ciascun passo successivo dell'algoritmo sostituisce o due punti assegnati oppure un punto assegnato e un punto sostituto oppure due punti sostituti con un altro punto sostituto, fino a quando il sottoinsieme non sia ridotto a tre punti.

Una volta trovato il punto di Steiner per questi tre punti, l'algoritmo procede a ritroso, tentando di determinare il punto di Steiner corrispondente a ciascun punto sostituto (*si veda l'illustrazione nella pagina a fronte in basso*). Un tentativo può non riuscire perché vi sono vincoli contraddittori sulla collocazione dei punti di Steiner. Se un tentativo riesce, tuttavia, si ottiene un albero di Steiner che congiunge con un ramo ciascun punto assegnato del sottoinsieme. Dopo aver considerato tutte le successioni di punti sostituti, l'algoritmo sceglie il più corto degli alberi di Steiner relativi al sottoinsieme. Combinando in tutti i modi possibili gli alberi di Steiner dei sottoinsiemi che risultano minimi, in modo che l'insieme originario dei punti assegnati sia coperto, si ottengono tutti gli alberi di Steiner localmente minimi ed è possibile determinare la geometria della rete globale minima.

L'algoritmo di Melzak può richiedere un tempo lunghissimo anche per problemi modesti perché considera un numero enorme di possibilità. Per esempio, a un problema con 10 punti corrispondono 512 sottoinsiemi di punti assegnati. Benché i sottoinsiemi di due punti non richiedano indubbiamente molta fatica, ciascuno dei 45 sottoinsiemi di otto punti ha due milioni di successioni di punti sostituti. Inoltre vi sono più di 18 000 modi di ricombinare gli alberi partendo dai sottoinsiemi.

I ricercatori hanno ovviamente scoperto metodi più efficaci per organizzare il calcolo e per aumentare la velocità dell'algoritmo. Invece di considerare la geometria del problema, questi metodi si concentrano sulle possibili configurazioni dei collegamenti della rete, cioè quella che si chiama la topologia della rete. Una topologia specifica quali punti sono collegati tra loro, ma non le posizioni effettive dei punti di Steiner. Supponendo assegnata una certa topologia, è possibile trovare con relativa rapidità un'opportuna successione di punti sostituti. Questo metodo per affrontare il problema accresce notevolmente la velocità di calcolo degli alberi di Steiner più brevi per i sottoinsiemi. Per un sottoinsieme di otto punti, per esempio, l'algoritmo deve considerare soltanto circa 10 000 topologie diverse anziché due milioni di successioni diverse di punti sostituti.

Dato che il numero delle topologie cresce rapidamente con il numero di punti del sottoinsieme, i problemi di Steiner potrebbero diventare più trattabili se fosse possibile limitarsi a considerare soltanto sottoinsiemi molto piccoli dell'insieme dei punti assegnati. Esperimenti compiuti con l'algoritmo di Melzak indicano che la rete di lunghezza minima per più di sei punti scelti a caso può di norma essere scomposta in reti di lunghezza minima per insiemi di punti più piccoli. Tuttavia, considerando particolari distribuzioni di punti chiamate scale, Fan R. K. Chung della Bell Communications Research e uno di noi (Graham) hanno dimostrato che esistono insiemi di punti arbitrariamente grandi per i quali non è possibile «scomporre» un albero di Steiner minimo. Una scala è una distribuzione in cui tutti i punti assegnati sono collocati a intervalli uguali lungo due rette parallele. Per questo particolarissimo problema di Steiner è stata scoperta una soluzione generale. Essa dimostra che il numero di punti di Steiner in un albero di Steiner di lunghezza minima per una scala con un numero dispari di «pioli» è quello massimo: il numero dei punti assegnati meno due. Un albero di Steiner siffatto non può essere scomposto perché la collocazione di ciascun punto di Steiner richiede che tutti i punti assegnati siano considerati simultaneamente. Quindi non si può sperare in una limitazione superiore del numero di punti contenuti nei sottoinsiemi dell'algoritmo di Melzak.

Molti ricercatori hanno poi migliorato l'algoritmo di Melzak trovando metodi più scaltri per ridurre la quantità di lavoro (*si veda l'illustrazione nella pagina seguente*). Questi metodi sfoltiscono,

La rete di lunghezza minima per tre punti *A*, *B* e *C* può essere ottenuta come segue. Costruito il triangolo equilatero *ACX* (*in verde*) sul lato più lungo del triangolo *ABC*, lo si inscrive in un cerchio (*in giallo*). Il punto *P* di intersezione del cerchio con il segmento che congiunge *B* al terzo vertice *X* del triangolo equilatero è detto punto di Steiner. Congiungendo con *P* i punti *A*, *B* e *C* si ottengono tre angoli di 120 gradi e si ricava la rete di lunghezza minima. La lunghezza del segmento *BX* risulta uguale alla lunghezza della rete.

L'algoritmo di Melzak riduce il problema della rete minima a problemi più semplici. *A* è il punto opportuno per scomporre il problema in uno con tre punti e in uno con cinque. Per costruire possibili alberi di Steiner per il problema con cinque punti si può sostituire una coppia di punti (per esempio *B* e *C*) con un unico punto (*X* in questo caso) costruendo un triangolo equilatero da una banda di *B* e *C*. Il problema è così ridotto a quattro punti: *X*, *D*, *G* e *A*. Ora si può sostituire una coppia di questi punti; in questo caso prima *D* e *X* con *Y* e poi *G* e *A* con *Z*. Ciascuno dei triangoli equilateri risultanti (*XDY*, *AGZ* e *BCX*) viene inscritto in un cerchio. I punti in cui il segmento *YZ* interseca due dei cerchi costituiscono i punti di Steiner *Q* e *R* e l'intersezione del segmento *XQ* con l'altro cerchio individua il punto di Steiner *P*. Poiché non si può stabilire in anticipo la migliore partizione e il miglior accoppiamento, per trovare l'albero minimo si debbono esaminare tutte le possibilità.

cioè eliminano, fasi dell'elaborazione che fornirebbero soltanto reti di grande lunghezza. Le nuove tecniche di sfoltimento hanno permesso di ottenere riduzioni sostanziali dei tempi di calcolo. Certi programmi basati sull'algoritmo di Melzak, come quello compilato nel 1969 da Ernest J. Cockayne dell'Università di Victoria, sono riusciti a risolvere tutti i problemi con nove punti e anche alcuni problemi con dodici punti in mezz'ora circa. Un programma scritto di recente da Cockayne e da Denton E. Hewgill, suo collega all'Università di Victoria, sfruttando un'efficace tecnica di sfoltimento introdotta da Pawel Winter dell'Università di Copenaghen risolve in pochi minuti tutti i problemi con 17 punti e la maggior parte dei problemi con 30 punti generati casualmente. Il metodo di sfoltimento di Winter riesce a eliminare possibili topologie con tale efficienza che attualmente la parte preponderante del tempo di calcolo è assorbita dalla fase di ricombinazione delle soluzioni per i sottoinsiemi.

Per ciascuno di questi programmi, tuttavia, il tempo di esecuzione può dipendere in maniera significativa sia dalla disposizione geometrica, sia dal numero dei punti. Inoltre il tempo di calcolo anche dell'algoritmo migliore cresce esponenzialmente con il numero dei punti e problemi di Steiner con 100 punti sono ancora ben lontani dalla nostra portata. Sarà mai possibile trovare un algoritmo efficiente in grado di calcolare le soluzioni dei problemi di Steiner di grandi dimensioni?

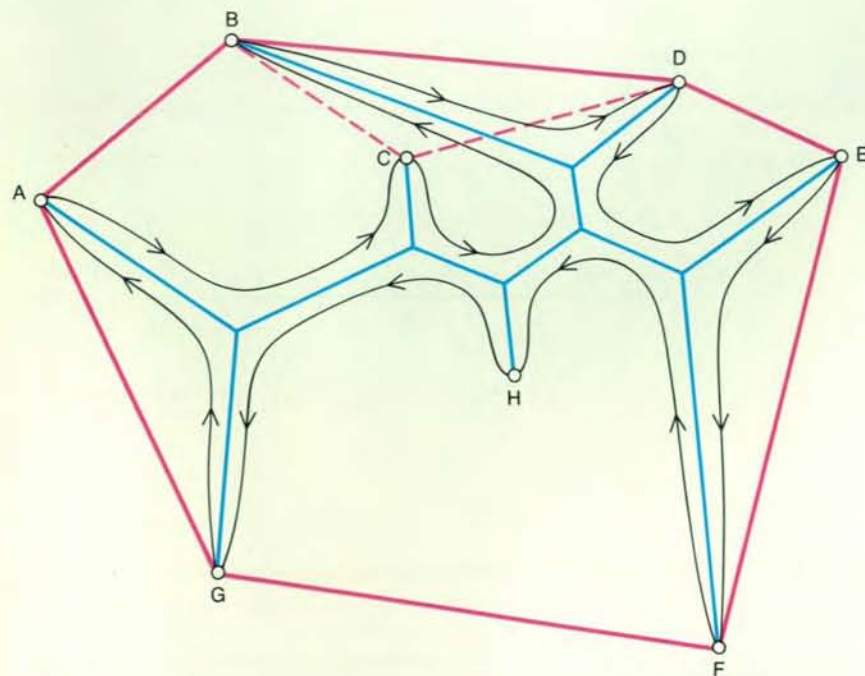
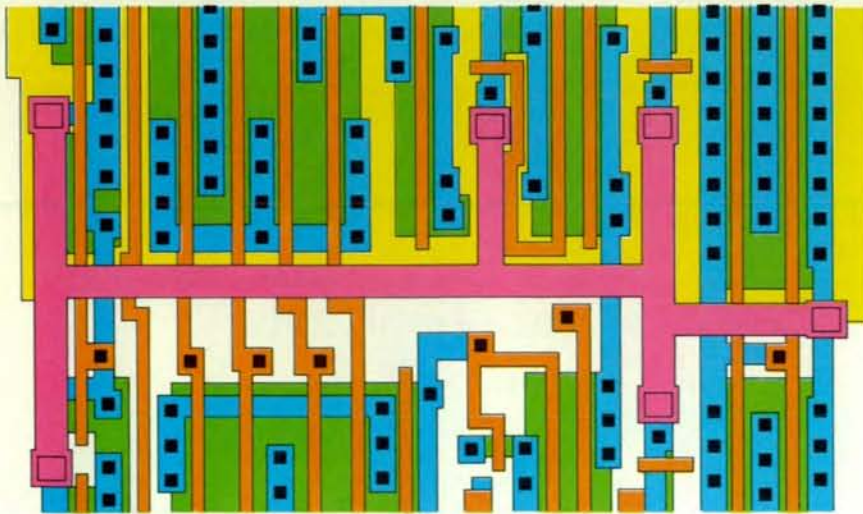
I progressi della teoria del calcolo hanno convinto la maggior parte dei ricercatori che gli algoritmi per i problemi di Steiner oggi esistenti non potranno subire miglioramenti sostanziali. In questa teoria viene assegnata una dimensione a ciascun esempio o caso particolare di un problema. Per i problemi di Steiner esiste una misura naturale della dimensione: il numero dei punti assegnati. Si considera quindi il numero d'operazioni elementari (quali per esempio addizioni, sottrazioni, confronti) che un algoritmo richiede per risolvere un caso particolare di una determinata dimensione. Dato che casi particolari diversi aventi la stessa dimensione possono richiedere un numero di operazioni diverso, si considera il numero massimo di operazioni in funzione della dimensione. Se il numero di operazioni cresce come la dimensione n del caso trattato elevata a qualche potenza, come nelle espressioni n^2, $5n$ o $6n+n^{10}$, si dice che la procedura è un algoritmo (in tempo) polinomiale. Questi algoritmi sono considerati efficienti, almeno sotto il profilo teorico. Se il numero delle operazioni cresce esponenzialmente con n, come nei casi 2^n, 5^n o $3n^2 \times 4^n$, allora si dice che la procedura è un algoritmo (in tempo) esponenziale.

Sebbene per problemi molto piccoli entrambi i tipi di algoritmi, in tempo polinomiale e in tempo esponenziale, siano praticabili, per problemi «grandi», i tempi di calcolo richiesti dagli algoritmi esponenziali sono così lunghi da essere assolutamente impraticabili. Per problemi abbastanza grandi, un algoritmo in tempo polinomiale fatto girare anche su macchina estremamente lenta fornisce la risposta prima di un algoritmo in tempo esponenziale fatto girare su un supercalcolatore.

Benché per il problema di Steiner siano stati trovati algoritmi esponenziali (per esempio l'algoritmo di Melzak), finora non è stato trovato alcun algoritmo polinomiale e le prospettive in questo senso non sono molto buone.

Nel 1971 Stephen A. Cook dell'Università di Toronto dimostrò che se si potesse trovare un algoritmo in tempo polinomiale anche per un solo problema appartenente alla classe dei cosiddetti problemi *NP*, (non polinomiali) quell'algoritmo potrebbe essere usato per risolvere in modo efficiente tutti gli altri problemi appartenenti a un'ampia classe di problemi difficili, compresi quelli *NP*. In seguito, uno di noi (Graham), in collaborazione con Michael R. Garey e David S. Johnson degli AT&T Bell Laboratories, dimostrò che il problema di Steiner è un problema *NP*. Siccome tutti i problemi *NP* hanno finora resistito ai tentativi di risoluzione di migliaia di ricercatori, si ritiene improbabile che un qualunque problema *NP*, incluso il problema di Steiner, possa essere risolto con un algoritmo in tempo polinomiale. Tuttavia dimostrare che i problemi *NP* non possono essere risolti in modo efficiente è oggi il problema più importante dell'informatica teorica.

Benché non appaia probabile che si possa trovare un algoritmo polinomiale efficiente per calcolare le reti di lunghezza minima, esistono algoritmi praticabili che permettono di individuare reti un poco più lunghe. Un chiaro esempio viene fornito dall'algoritmo compilato per risolvere il problema dell'albero ricoprente minimo, con cui si cerca la rete di segmenti più corta che colleghi un insieme di punti assegnati senza aggiungerne di nuovi. Per risolvere questo tipo di problema, si collegano i due punti assegnati tra loro più vicini e in ciascuno dei passi successivi si collegano le coppie di punti più vicine che possono essere collegate senza che si formi alcun cammino chiuso. Infatti da un cammino chiuso è possibile rimuovere un ramo lasciando i punti assegnati ancora collegati dai rami restanti.

Edgar N. Gilbert e Henry O. Pollak dei Bell Laboratories hanno avanzato l'ipotesi che il rapporto tra un albero di Steiner di lunghezza minima e un albero ricoprente minimo sia almeno $\sqrt{3}/2$, cioè che la lunghezza dell'albero di Steiner sia al massimo del 13,4 per cento

I metodi di sfoltimento accrescono l'efficienza degli algoritmi per la ricerca delle reti più corte. Un modo per sfoltire possibili reti è quello (ideato da Cockayne) di considerare l'ordine in cui un elastico (*in rosso*) teso intorno all'insieme dei punti assegnati, li tocca. L'elastico tocca tutti i punti tranne *C* e *H*, ma *C* può essere inserito nella successione perché l'angolo formato da *C* e da due punti consecutivi in contatto con l'elastico misura più di 120 gradi. L'ordine dei punti è allora *ABCDEFG*. Un cammino ininterrotto (*in nero*) tracciato intorno a una possibile rete (*in blu*) tocca i punti nell'ordine *ACBDEFHG*. Dato che *B* e *C* sono invertiti rispetto all'ordine dato dall'elastico, questa rete può essere sfoltita.

Certe varianti del problema della rete di lunghezza minima sono state applicate alla progettazione di circuiti elettronici per aumentarne la velocità di funzionamento. Qui la rete minima di conduttori orizzontali e verticali che collega un insieme di terminali è in rosso. Gli altri colori corrispondono a conduttori e terminali disposti su strati più profondi.

circa inferiore alla lunghezza dell'albero ricoprente minimo. Il rapporto $\sqrt{3}/2$ si presenta in un semplice esempio: quando i tre punti assegnati formano un triangolo equilatero. Sebbene la congettura non sia stata ancora dimostrata, Chung e uno di noi (Graham) hanno dimostrato che l'albero di Steiner è al massimo del 17,6 per cento più corto dell'albero ricoprente minimo.

Un albero ricoprente minimo può essere spesso accorciato del 3 o 4 per cento aggiungendo in modo opportuno alcuni punti di Steiner e riaggiustando l'albero. Uno di noi (Bern) ha dimostrato che algoritmi inesatti di questo genere hanno qualche giustificazione teorica, dal momento che la lunghezza media dell'albero riaggiustato sarà un po' minore della lunghezza media di un albero ricoprente minimo.

I problemi dell'albero ricoprente minimo e della rete di lunghezza minima sono stati applicati alla costruzione di reti di tubature, telefoniche e stradali. Le soluzioni, approssimate o esatte, possono fornire indicazioni sulla configurazione delle reti e sulle quantità di materiali necessarie. Versioni più complicate del problema di Steiner possono tener conto della necessità di evitare certi ostacoli geografici o di trovare i collegamenti più brevi nel quadro di reti già esistenti.

Forse l'applicazione più frequente del problema di Steiner si ha nella progettazione di circuiti elettronici. Se la rete di conduttori di un circuito integrato è corta, per caricarla e scaricarla occorre un tempo inferiore rispetto a una rete lunga e quindi la velocità di funzionamento del circuito aumenta. Nel caso dei circuiti tuttavia il problema della rete di lunghezza minima è soggetto a un ulteriore, differente vincolo geometrico, poiché i conduttori di un circuito sono in genere disposti secondo due sole direzioni, verticale e orizzontale.

Questo problema, che ha preso il nome di problema di Steiner rettilineo, fu studiato per la prima volta nel 1965 da Maurice Hanan del Thomas J. Watson Research Center della IBM a Yorktown Heights, presso New York. Come nel problema di Steiner originale, anche nel caso rettilineo la soluzione è un albero contenente i punti di Steiner e i punti assegnati, ma i rami s'incontrano a 90 o a 180 gradi. Sebbene nel problema rettilineo i punti di Steiner possano in linea di principio trovarsi ovunque, come nel problema originale, Hanan dimostrò che in una rete rettilinea di lunghezza minima la collocazione dei punti di Steiner può essere soggetta a limitazioni. Per ogni punto assegnato vengono tracciate una retta orizzontale e una verticale; ogni intersezione di due rette individua allora un possibile punto di Steiner. Un algoritmo può prendere in esame tutti i sottoinsiemi di possibili punti di Steiner allo scopo di calcolare la rete minima. Al crescere del numero dei punti assegnati, tuttavia, il tempo richiesto da un algoritmo così rozzo per giungere alla soluzione cresce esponenzialmente. Algoritmi più raffinati, ma ugualmente esponenziali, possono risolvere problemi di Steiner rettilinei con circa 40 punti.

Una versione rettilinea del problema dell'albero ricoprente minimo può essere risolta in modo efficiente dall'algoritmo che sceglie a ogni passo il collegamento più breve, a meno che quel collegamento non formi un cammino chiuso. Frank K. Hwang dei Bell Laboratories ha dimostrato che la lunghezza di un albero di Steiner rettilineo non è mai inferiore di più di un terzo a quella di un albero ricoprente rettilineo minimo.

L'applicazione più sorprendente del problema di Steiner si ha nel campo della filogenesi. David Sankoff dell'Università di Montreal e altri ricercatori hanno messo a punto una versione del problema di Steiner che viene impiegata per calcolare alberi filogenetici plausibili. I ricercatori isolano dapprima una particolare proteina che è comune agli organismi che vogliono classificare. Per ciascun organismo determinano poi la sequenza degli amminoacidi che formano la proteina e individuano un punto in una posizione data dal numero di differenze tra la proteina dell'organismo e la proteina corrispondente di altri organismi. Gli organismi con sequenze simili sono quindi definiti vicini tra loro e gli organismi con sequenze dissimili sono definiti lontani. In una rete di lunghezza minima per questa collocazione astratta di punti assegnati, i punti di Steiner corrispondono agli antenati più plausibili e i rami corrispondono a una relazione tra organismo e antenato che presuppone il minimo numero di mutazioni. Tuttavia, dato che il problema di Steiner filogenetico non è certo più facile degli altri problemi di Steiner, esso è servito, tranne quando è applicato a pochi organismi, più per esperimenti concettuali che come strumento pratico di ricerca.

Benché negli ultimi anni la nostra conoscenza degli algoritmi abbia fatto registrare un gran numero di progressi, l'affascinante problema della rete di lunghezza minima resta ancora troppo difficile per venire completamente padroneggiato. Il problema può essere enunciato in termini semplici, eppure le sue soluzioni sfidano le nostre capacità analitiche. Per quanto insignificante possa sembrare, una variazione anche esigua della geometria di un problema può tuttavia modificare radicalmente la corrispondente rete minima e questa sensibilità rende assai stimolanti anche questioni marginali relative alle reti di lunghezza minima. Indubbiamente il problema della rete minima continuerà a resisterci e ad affascinarci ancora per molti anni.

BIBLIOGRAFIA

GILBERT E. N. e POLLAK H. O., *Steiner Minimal Trees* in «SIAM Journal on Applied Mathematics», 16, n. 1, gennaio 1968.

MELZAK Z. A., *Companion to Concrete Mathematics*, John Wiley & Sons, Inc., 1973.

WINTER PAWEL, *An Algorithm for the Steiner Problem in the Euclidean Plane* in «Networks», 15, n. 3, autunno 1985.

WINTER PAWEL, *Steiner Problem in Networks: A Survey* in «Networks», 17, n. 2, estate 1987.

André-Marie Ampère

Fu il primo ricercatore a valutare quantitativamente gli effetti magnetici della corrente elettrica, ma si interessò anche attivamente di filosofia della scienza, rivendicando l'importanza delle entità teoriche non osservabili

di L. Pearce Williams

Quando una matricola mi chiede chi fu André-Marie Ampère, mi accade spesso di rispondere: «Probabilmente il suo nome vi è più noto nella forma minuscola "ampere", che si ha spesso occasione di usare quando si comprano dei comuni fusibili». Ampère è noto soprattutto come il fondatore della scienza dell'elettrodinamica. All'inizio dell'Ottocento eseguì le prime ricerche sistematiche sui campi magnetici prodotti da correnti elettriche, scoprì e valutò quantitativamente le forze che agiscono fra conduttori elettrici percorsi da corrente e fu il primo ricercatore ad avanzare l'ipotesi che il magnetismo che si osserva nei magneti permanenti sia causato da minuscole correnti elettriche circolanti all'interno delle molecole del materiale magnetico. I risultati da lui conseguiti sono ricordati oggi dal nome dell'unità di misura dell'intensità della corrente elettrica nel Sistema internazionale: l'ampere (simbolo: A), l'unità che viene usata per specificare i diversi valori di corrente per i quali sono tarati i comuni fusibili.

Ampère è meno noto per i risultati conseguiti come filosofo della scienza, anche se essi furono sotto certi aspetti altrettanto significativi. Egli fu il primo scienziato di rilievo ad adattare le concezioni del filosofo tedesco Immanuel Kant perché fornissero un fondamento filosofico allo studio della fisica e della chimica. Ampère prese l'avvio dalla teoria della conoscenza di Kant - la teoria di ciò che gli esseri umani possono conoscere e di come possono pervenire a conoscerlo - nel suo tentativo di sviluppare un metodo pratico per l'indagine scientifica. La filosofia kantiana finì col pervadere le scienze fisiche, particolarmente verso la fine dell'Ottocento e all'inizio del Novecento, e il metodo di indagine di Ampère, che lo guidò nelle sue ricerche di elettrodinamica, sopravvive oggi in forma modificata come una delle metodologie scientifiche più comunemente accettate.

Ampère, nato nel 1775, trascorse l'infanzia nel villaggio di Poleymieux-les-Mont-d'Or, nei pressi di Lione. Si istruì leggendo avidamente i libri della biblioteca paterna e volumi che comprava a Lione. Divorò l'*Encyclopédie* di Denis Diderot e Jean-Baptiste d'Alembert, che era stata appena completata, fissando nella mente intere voci grazie alla sua memoria fotografica. Era particolarmente interessato alla matematica; lesse con grande interesse i trattati del matematico svizzero Leonhard Euler sull'algebra avanzata, sulla teoria delle probabilità e sul calcolo infinitesimale e verso i vent'anni padroneggiava l'opera fondamentale di Joseph-Louis Lagrange *Mécanique analitique*. I suoi interessi al di fuori della matematica furono letteralmente enciclopedici. Ampère studiò le opere di scienza naturale del celebre Georges-Louis Leclerc, conte di Buffon, imparò il greco, il latino e l'italiano, cercò di sviluppare un linguaggio universale che fosse fondato sui principi più aggiornati della linguistica, si dedicò alla letteratura francese, scrisse poesie, studiò botanica e sviluppò metodi per rendere sistematiche le sue osservazioni del mondo naturale. Questa vastità di interessi fu una caratteristica costante di tutta la sua vita.

Fino ai 18 anni Ampère, figlio unico in una famiglia che disponeva di mezzi considerevoli, condusse un'esistenza quasi idilliaca, vagando ovunque desiderasse - fisicamente e intellettualmente - in un mondo sicuro e stabile. Nel 1793, però, cominciò una serie di tragedie personali che costellarono il resto della sua esistenza. In quello stesso anno suo padre fu ghigliottinato come controrivoluzionario. Nel 1803, dopo quattro anni di felice matrimonio, morì la sua prima moglie (madre di suo figlio Jean-Jacques) e dopo altri quattro anni egli si risposò; questa seconda unione fu però disastrosa e si concluse con il divorzio, lasciandogli una bambina piccola. Dopo alcuni anni di relativa tranquillità, nel 1819 il figlio Jean-Jacques rinunciò a ogni seria prospettiva per unirsi all'*entourage* di Madame Récamier, la grande bellezza dell'epoca, che faceva collezione di ammiratori, tenendoli però risolutamente a distanza. Jean-Jacques la frequentò per i successivi vent'anni, ignorando le preghiere del padre di tornare a casa e di abbracciare una qualche carriera seria. Nel 1827 Ampère diede sua figlia in sposa a un ex ufficiale dell'esercito napoleonico, accorgendosi però solo in seguito che il genero era alcolizzato e malato di mente.

Per tutta la vita Ampère soffrì di un costante deterioramento delle condizioni di salute, che ostacolò progressivamente ogni suo impegno scientifico. Fu anche tormentato, per la maggior parte della sua vita, da una situazione di insicurezza finanziaria che lo costrinse ad accettare una quantità di lavori mal pagati, la maggior parte dei quali comportava, a un qualche livello, l'insegnamento della matematica.

La carriera scientifica di Ampère fu ostacolata non solo da queste circostanze personali, ma anche dalla vastità stessa dei suoi interessi intellettuali. In una serie di lettere da lui scritte all'età di vent'anni e venute in luce solo di recente lo vediamo passare dalla meccanica teorica alla costruzione pratica di macchine, alla teoria del volo degli aquiloni, ai linguaggi artificiali, alla musica, all'astronomia, alla botanica, ai sistemi di classificazione. Non riuscì mai a concentrarsi in modo esclusivo su un'unica disciplina e ad applicare in essa tutto il suo genio. Persino mentre stava sviluppando febbrilmente la sua teoria dell'elettrodinamica, continuava a speculare su problemi di metafisica e di filosofia. Ampère non vide mai alcuna contraddizione in tutto questo, essendo convinto dell'esistenza di un'unità fondamentale alla base di ogni conoscenza. Come vedremo più avanti, la sua ultima opera importante fu proprio un tentativo di dimostrare questo punto di vista.

André-Marie Ampère (*a sinistra*) e François Arago esaminano gli effetti magnetici della corrente elettrica in una ricostruzione immaginaria e un po' imprecisa (risalente alla fine dell'Ottocento) delle ricerche che eseguirono assieme. Ampère, che fu uno dei creatori dell'elettrodinamica, cercò di fondare su considerazioni filosofiche un quadro metodologico in cui inserire la ricerca scientifica. Arago riferì per primo ad Ampère e agli altri membri dell'Académie Française des Sciences la scoperta di Hans Christian Oersted che una corrente elettrica fa deviare l'orientazione dell'ago di una bussola posta vicino al conduttore percorso da corrente.

Quando Ampère iniziò le sue ricerche scientifiche, la filosofia francese era dominata da una scuola i cui membri erano stati sprezzantemente battezzati da Napoleone «ideologi» (*idéologues*). Essi sostenevano di avere determinato le regole complete per una metodologia scientifica appropriata. Nella loro concezione, la mente umana è un recettore passivo di impressioni sensoriali. Sulla base di queste impressioni la mente crea una serie di immagini (comprendenti sensazioni dell'olfatto, dell'udito, del gusto e del tatto, oltre che della vista) che rappresentano il mondo esterno. La mente è in grado di richiamare immagini e di paragonarle, per determinarne le differenze e accertare in che modo la successione delle immagini muti nel tempo. Ogni regolarità osservata può allora servire da base per leggi scientifiche. Tuttavia, non c'è alcun modo per sapere se esista realmente un mondo esterno. L'unica realtà nota risiede nelle immagini.

In questa concezione del mondo non esiste causalità, ma solo una successione di immagini. È dunque impossibile, in un certo senso, spiegare i fenomeni, ossia descrivere le forze fisiche per mezzo delle quali una certa causa determina un particolare effetto. Ne deriva che non possono esistere teorie scientifiche intese in un senso moderno del termine. Da un lato, la scienza diventa un mero procedimento tassonomico: lo scienziato ordina sensazioni simili in gruppi organizzati in modo appropriato. Dall'altro, essa diventa positivistica: lo scienziato esprime in leggi matematicamente rigorose le regolarità osservate. Un buon esempio del primo tipo di scienza ci viene offerto dal sistema chimico di Antoine Lavoisier, che si fondava su una descrizione esatta, su una nomenclatura precisa e su una appropriata classificazione di elementi e composti chimici. Il secondo tipo di scienza è rappresentato dagli studi di Jean-Baptiste Fourier sul calore, nei quali egli ignorò del tutto le cause del calore concentrandosi invece sulla determinazione delle leggi matematiche che regolano la propagazione termica. Una impostazione filosofica di questo tipo è adeguata per molte discipline scientifiche, ma esclude virtualmente campi come la fisica dei fenomeni microscopici, la quale si fonda su entità teoriche che non sono suscettibili di osservazione, come atomi e molecole.

Le ricerche scientifiche giovanili di Ampère non furono limitate da questa impostazione filosofica; egli iniziò la sua carriera come matematico e la matematica non deve riferirsi necessariamente a una realtà fisica esterna. Il primo saggio edito di Ampère, scritto nel 1802 per farsi conoscere e poter puntare alla cattedra di matematica al Liceo napoleonico di Lione, era sulla teoria matematica del gioco d'azzardo. In questo breve trattato Ampère dimostrò che un giocatore che, disponendo di una quantità finita di denaro, si trovi a giocare o contro un singolo avversario che disponga di risorse finanziarie infinite o contro un gran numero di avversari con risorse finite perderà necessariamente tutto in capo a un tempo finito. Ampère scrisse anche saggi di meccanica teorica e la sua opera matematica più vasta - redatta in vista della sua nomina a membro dell'Académie Française des Sciences - fu un trattato sulle equazioni differenziali alle derivate parziali. Vari altri saggi matematici completano questa fase della sua carriera. Se il suo impegno fosse stato limitato alla matematica, oggi egli sarebbe noto al massimo come un matematico competente e a volte innovatore che fu tuttavia oscurato da contemporanei della grandezza di Laplace, Poisson, Cauchy e Fourier.

Una delle ragioni per cui Ampère non riuscì ad aprire vie nuove in matematica fu che nel 1805 questa disciplina aveva cominciato ad annoiarlo. Aveva trovato due nuove passioni: la metafisica e la chimica.

Il vivo interesse di Ampère per la metafisica trasse origine dalla sua iniziazione alla filosofia degli ideologi nel 1804, quando, dopo la morte della prima moglie, egli si trasferì da Lione a Parigi. Ampère entrò a far parte di un piccolo gruppo di filosofi che si riunivano nel villaggio suburbano di Auteuil per discutere l'opera dell'abate Condillac, il fondatore della scuola degli ideologi. Egli fu ben presto contrariato dalle concezioni degli ideologi in quanto essi escludevano l'esistenza di Dio e dell'anima immortale e abbandonò questo gruppo alla ricerca di un'alternativa. Trovò allora ispirazione per la sua filosofia negli scritti di Kant.

Il filosofo tedesco divideva il mondo in due ambiti: il regno dei fenomeni e quello dei noumeni. I fenomeni sono eventi quali sono percepiti dalla mente umana: sono cioè sensazioni. I noumeni sono le cause dei fenomeni, le cosiddette «cose in sé», gli oggetti quali esistono nella realtà. Gli esseri umani non potranno mai conoscere direttamente i noumeni: essi sono le fonti dei segnali

L'idea che le molecole avessero la forma di poliedri formò la base concettuale delle ricerche di chimica teorica compiute da Ampère fra il 1808 e il 1815. Motivato dall'asserzione di Immanuel Kant che la vera scienza dovrebbe fondarsi su principi matematici, Ampère tentò di spiegare la chimica in termini geometrici. Secondo l'ipotesi di Ampère, le molecole elementari dovevano essere costituite da atomi puntiformi disposti come i vertici dei solidi regolari o quasi regolari illustrati nella fila superiore di questa tavola (tratta da un articolo di Ampère pubblicato negli «Annales de Chimie»). Fra queste molecole erano possibili reazioni solo nel caso che queste potessero produrre solidi dotati di un certo grado di regolarità e di simmetria, come quelli illustrati nella fila centrale e in quella inferiore della tavola.

Ampère verificò in vari esperimenti l'ipotesi fondamentale che ogni forma di magnetismo (anche quello dei magneti permanenti) sia causata da correnti elettriche che fluiscono circolarmente. Se l'ipotesi fosse vera, solenoidi percorsi da correnti che hanno la stessa direzione (*a*) dovrebbero respingersi fra loro come magneti permanenti con i poli nord orientati nella stessa direzione. L'esperimento rivelò invece che i solenoidi si attraevano. Quando Ampère provò a ripeterlo con spirali piane, queste si comportarono come i magneti (*b*). Ampère notò che i tratti rettilinei di conduttore vicini alla batteria, percorsi da corrente nello stesso verso (*a destra*), si attraevano reciprocamente; ne dedusse che, poiché i solenoidi avevano le spire molto distanziate fra loro, il flusso circolare di corrente in essi era insignificante rispetto a quello longitudinale, cosicché i solenoidi si attraevano reciprocamente come fossero stati fili rettilinei. Ampère verificò l'ipotesi avvolgendo i solenoidi attorno a tubi di vetro e facendo ripassare il filo all'interno di ciascun tubo, in modo che la corrente di ritorno cancellasse gli effetti della corrente che fluiva in senso longitudinale. I solenoidi così avvolti, come si vede nella versione *c* dell'esperimento, si respingevano.

che agiscono sui nostri sensi e noi possiamo percepire solo i segnali, non le fonti. Secondo Kant, quindi, noi non potremo mai conoscere realmente qualcosa di definito sui noumeni.

Fu su questo punto che Ampère si staccò da Kant. Egli notò che si osservano spesso relazioni, o rapporti (*rapports*) per usare il suo termine, fra i fenomeni. Quei rapporti, secondo Ampère, dovevano essere analoghi ai rapporti fra i noumeni che stanno alla base dei fenomeni. Dovrebbe perciò sussistere la possibilità di imparare qualcosa sulle interazioni fra i noumeni inosservabili studiando i rapporti fra fenomeni.

Questa asserzione filosofica fu il fondamento su cui Ampère costruì il suo metodo di indagine scientifica, che spiegò in una lettera scritta nel 1810 al vecchio amico Marie-François Pierre Maine de Biran. Secondo il metodo di Ampère, lo scienziato spiega i fenomeni ipotizzando l'esistenza di certe entità noumeniche, dopo di che si impegna in un processo di deduzione: accettando l'esistenza delle entità teoriche, quali nuovi risultati sperimentali - fenomeni - ci si possono attendere? La deduzione viene poi verificata per mezzo dell'esperimento. Si può assumere l'esistenza di entità teoriche nella misura in cui esse forniscono una spiegazione dei fenomeni verificabile sperimentalmente. La probabilità che la teoria sia vera aumenta quando viene dimostrata la sua capacità di resistere alla prova dell'esperimento. Il metodo di Ampère è chiamato oggi metodo ipotetico-deduttivo ed è considerato da molti il modo più appropriato per compiere ricerche scientifiche.

Un buon esempio dell'applicazione del metodo è l'analisi compiuta da Ampère della legge dei volumi o delle combinazioni semplici, scoperta nel 1808 da Joseph-Louis Gay-Lussac. Secondo questa legge fenomenica, i gas si combinano in rapporti volumetrici semplici; per esempio, quando si fanno reagire due litri di idrogeno con un litro di ossigeno, si producono esattamente due litri di vapore acqueo. Per gli ideologi non ci si poteva spingere oltre questo livello; non si poteva dire *perché* la legge fosse vera. Ampère era deciso a non fermarsi a quel punto. Nel 1814 sostenne, in uno dei primi saggi moderni di chimica teorica, che i fenomeni potevano essere spiegati solo supponendo che volumi uguali di gas diversi, nelle stesse condizioni di temperatura e di pressione, contenessero lo stesso numero di molecole. L'esistenza di molecole - noumeni inosservabili - è il fondamento su cui si basò Ampère per spiegare i fenomeni descritti da Gay-Lussac.

Subito dopo avere escogitato il suo metodo, Ampère lo applicò sviluppando un brillante quadro teorico per la chimica. Kant aveva sostenuto che la scienza deve fondarsi su principi matematici *a priori*: principi che egli riteneva intrinseci alla struttura della mente umana e non derivati dall'osservazione. Seguendo l'insegnamento di Kant, Ampère cercò di dedurre le leggi dell'affinità chimica (le leggi che determinano quali reazioni chimiche siano possibili) supponendo l'esistenza di molecole ipotetiche di natura geometrica.

Ampère suppose che ogni molecola fosse composta da atomi puntiformi disposti nello spazio come i vertici di un solido geometrico semplice, quale un tetraedro, un ottaedro o un cubo. Le uniche combinazioni chimiche che potevano verificarsi erano quelle che davano origine a solidi geometrici aventi un certo grado di simmetria e regolarità tridimensionali. Nel quadro teorico di Am-

père la sconcertante arbitrarietà dell'attività chimica poteva essere ridotta alla certezza matematica: la chimica poteva essere fondata sulla geometria, che secondo Kant era la forma più pura di matematica. Si deve osservare a questo punto che nessuno dei due saggi sulla chimica di Ampère trovò molto favore presso i chimici, che erano ostili sia alle sue speculazioni, sia alla sua formulazione matematica.

I risultati ottenuti da Ampère prima del 1819, quindi prima di compiere 44 anni, non meriterebbero probabilmente più di una nota a piè di pagina nella storia della fisica. Diversamente dai suoi contemporanei Augustin Fresnel (grande amico di Ampère e uno degli ideatori della teoria ondulatoria della luce) e Sadi Carnot (uno dei fondatori della termodinamica), che morirono rispettivamente a 39 e a 36 anni, Ampère conseguì i suoi massimi risultati scientifici in età matura, dopo aver disperato di poter lasciare un segno nella storia della scienza.

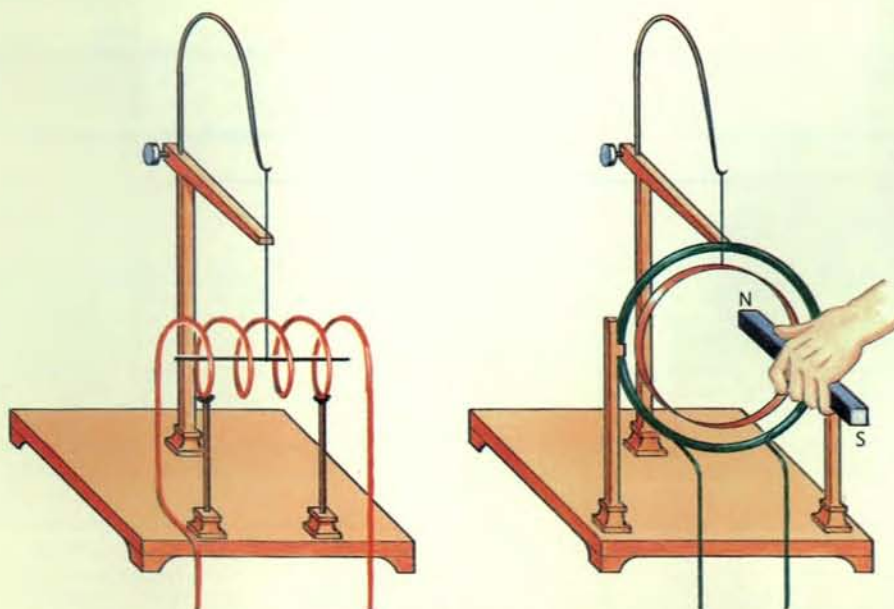
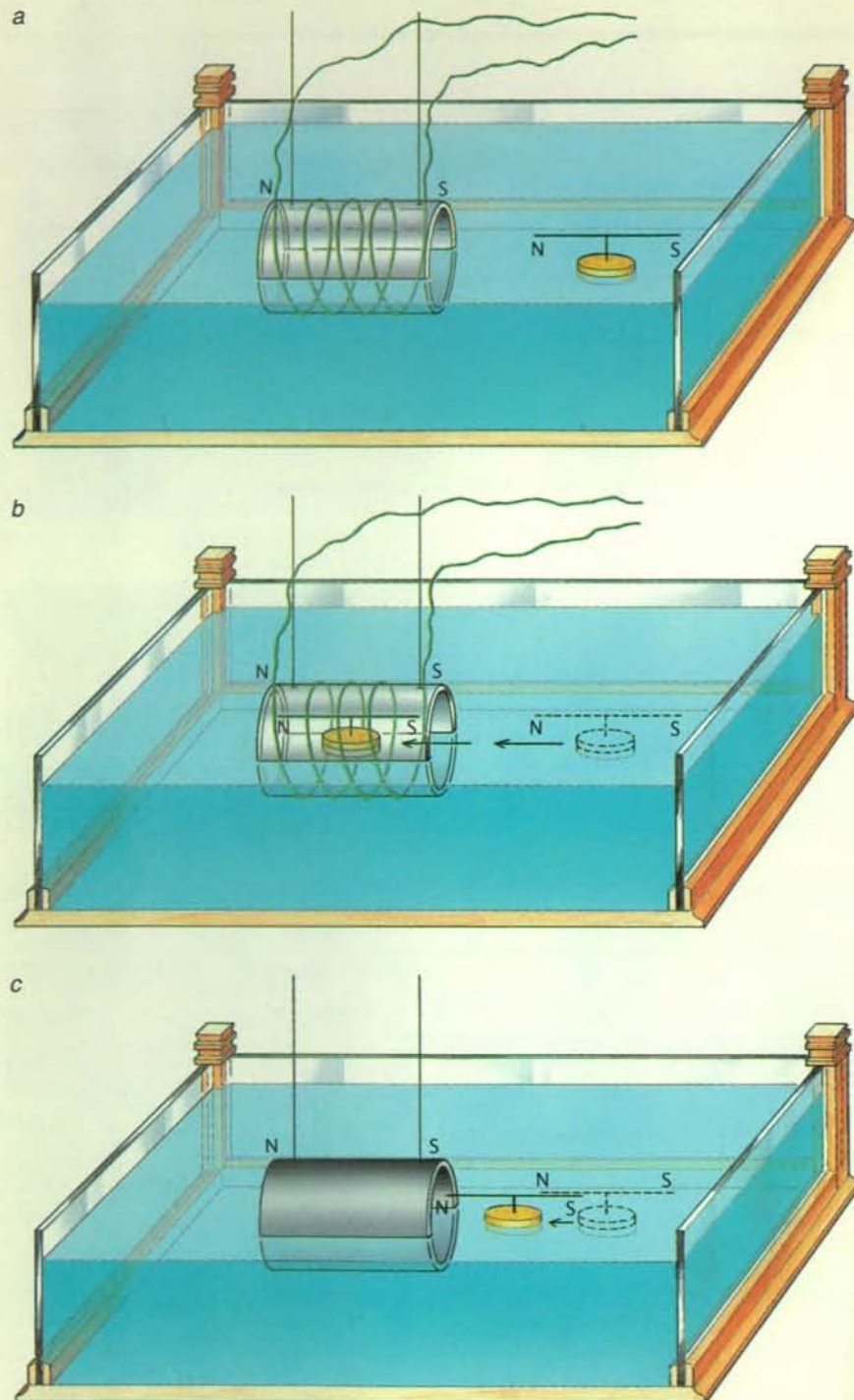
I primi esperimenti di Ampère in elettrodinamica implicarono l'uso della pila voltaica, che era stata inventata nel 1800 da Alessandro Volta. Una pila voltaica è una cella elettrochimica, molto simile a una moderna batteria per automobili. Se i poli di una pila voltaica vengono collegati fra loro da un conduttore elettrico, lungo il conduttore scorrerà una corrente elettrica, mentre le reazioni chimiche all'interno della cella opereranno in modo da conservare la differenza di potenziale fra i poli.

La corrente elettrica continua era un fenomeno nuovo all'inizio del XIX secolo e le prime teorie sulla corrente elettrica si basavano in gran parte sulle teorie dell'elettricità statica. Ben pochi prevedevano che la corrente avesse effetti magnetici, dato che nel penultimo decennio del Settecento Charles-Augustin de Coulomb aveva dimostrato che le forze associate all'elettricità statica sono distinte dal magnetismo. Soltanto alcuni «filosofi della natura» tedeschi o influenzati dalla cultura tedesca, che credevano nella fondamentale unità di tutte le forze, cercavano una relazione fra elettricità e magnetismo.

Uno di questi filosofi fu il danese Hans Christian Oersted. Nel 1807 e nel 1812 egli pubblicò opere nelle quali sostenne, per ragioni filosofiche, che elettricità e magnetismo dovevano essere correlati. Nell'inverno e nella primavera del 1820 egli scoprì finalmente una relazione fra elettricità e magnetismo tenendo l'ago di una bussola accanto a un lungo conduttore elettrico; quando nel conduttore passava corrente, l'ago veniva deviato. La scoperta di Oersted fu pubblicata su tutti i principali periodici scientifici del tempo.

La notizia fu portata a Parigi da un amico di Ampère, Jean-François-Dominique Arago, che aveva assistito a un esperimento di questo tipo durante una sua visita a Ginevra. I membri dell'Académie des Sciences ascoltarono la relazione di Arago con scetticismo e rimasero convinti solo quando egli ripetè l'esperimento dinanzi a loro, l'11 settembre di quell'anno. Alla dimostrazione era presente Ampère, che, tornato a casa, cominciò a indagare personalmente su quell'effetto. Egli si rese conto immediatamente che Oersted non aveva compreso a fondo il risultato dell'esperimento perché non aveva tenuto conto del magnetismo terrestre. L'entità della deviazione dell'ago della bussola dipendeva dall'angolo esistente fra il conduttore percorso dalla corrente e il campo magnetico terrestre.

Ampère si accinse immediatamente a identificare i veri effetti del passaggio di corrente elettrica sull'orientazione dell'ago di una bussola escogitando una disposizione di magneti liberamente rotanti che neutralizzassero il campo magnetico terrestre in una piccola regione dello spazio. Con sua grande soddisfazione trovò che in questo modo l'ago della bussola si disponeva sempre ad angolo retto rispetto al conduttore percorso dalla corrente. Si rese conto quindi che l'ago di una bussola poteva essere usato come parte di uno strumento per rilevare una corrente elettrica. Mediante il suo nuovo strumento, che chiamò galvanometro, riuscì a definire il percorso della corrente lungo un circuito formato da un conduttore elettrico e da una pila voltaica.

Fino ad allora si era supposto che, se si collegavano con un conduttore elettrico i due poli di una pila voltaica, i meccanismi operanti all'interno della pila fossero distinti dalla corrente che percorreva il filo. Ampère trovò, con sua sorpresa, che la corrente che passava attraverso la pila voltaica era la stessa che percorreva la restante parte del circuito. Che cosa sarebbe accaduto, si chiese, se avesse costruito una pila circolare, una pila foggiata ad anello in modo che il polo positivo fosse stato a contatto con il polo negativo? Da prove indirette pare che Ampère abbia costruito tale pila durante quello stesso fruttuoso mese di settembre del 1820, osservando che essa generava un campo magnetico simmetrico. A questo punto egli concepì l'ipotesi che avrebbe difeso per il resto della sua vita: il magnetismo non è altro che un complesso di correnti elettriche che si muo-

La natura delle correnti che sono alla base del magnetismo fu studiata in due esperimenti eseguiti separatamente da Arago e da Ampère. Nel primo (*a sinistra*) Arago realizzò un solenoide di filo di rame con un ago di ferro sospeso all'interno; facendo passare corrente attraverso il solenoide notò che l'ago si magnetizzava. Ampère pensò di verificare se le correnti circolari responsabili del magnetismo permanente fluissero attorno a ogni particella di materiale magnetico o attorno all'asse del magnete nel suo complesso. Nel suo esperimento (*a destra*), egli preparò una spira di filo di rame isolato (*in verde*) e sospese al suo interno un anello fatto di nastro di rame. Ampère pensò che se la corrente nel solenoide aveva magnetizzato l'ago facendo fluire correnti attorno al suo asse centrale essa doveva magnetizzare anche l'anello di rame, perché le correnti potevano circolare attorno al suo asse. Egli fece passare una corrente lungo la spira e tenne un magnete a barra in prossimità dell'anello per vedere se esso fosse diventato un magnete «permanente». Poiché ciò non avvenne, Ampère si convinse che le correnti nei magneti permanenti circolano attorno alle singole particelle e non attorno all'asse magnetico. (L'esperimento non è ineccepibile da un punto di vista moderno, ma la conclusione tratta da Ampère è perfettamente valida.)

Un'apparente smentita dell'ipotesi fondamentale di Ampère sul magnetismo venne da esperimenti compiuti da Michael Faraday. Faraday avvolse un filo di rame isolato attorno a un tubo di vetro, sospendendolo in un recipiente pieno d'acqua (*a*). Poi fissò un lungo ago magnetizzato su un pezzo di sughero galleggiante presso il tubo e fece passare corrente elettrica nel solenoide. Secondo Faraday, se il solenoide percorso da corrente fosse stato esattamente simile a un magnete, il sughero avrebbe dovuto muoversi verso il tubo fino a che l'ago si fosse trovato appena all'interno del tubo; un polo magnetico dell'ago sarebbe stato allora il più vicino possibile al polo opposto del solenoide-magnete. Invece (*b*) il sughero entrò completamente nel tubo finché entrambi i poli dell'ago vennero a trovarsi in prossimità dei poli corrispondenti del solenoide. In un altro esperimento (*c*), Faraday arrotolò un foglio d'acciaio a forma di tubo, lo magnetizzò e lo sospese nell'acqua. Questa volta l'ago si mosse verso un'estremità del tubo e si fermò. Faraday sostenne perciò che il solenoide percorso da corrente e il magnete non erano identici. Ampère riuscì a riconciliare questo risultato con la sua ipotesi sottolineando che nel solenoide l'ago si era trovato all'interno della corrente mentre nel tubo d'acciaio era rimasto all'esterno delle piccole spire di correnti molecolari; era perciò prevedibile un comportamento diverso nei due casi.

vono in circolo. Era giunto a questo punto il 18 settembre, quando lesse la sua prima relazione all'accademia. Nessuno però gli dedicò molta attenzione.

Fra il 18 e il 25 settembre, la data fissata per la successiva riunione dell'accademia, Ampère applicò il suo metodo. Dopo avere ipotizzato una spiegazione noumenica - correnti elettriche circolari - per il fenomeno del magnetismo, il passo seguente era quello di sottoporre a verifica sperimentale deduzioni fondate sull'ipotesi: dimostrare cioè che correnti elettriche circolari producono gli stessi effetti dei magneti permanenti. Egli tentò innanzitutto di dimostrare questo effetto utilizzando un conduttore elettrico di rame, avvolto in modo da formare un solenoide (una bobina costituita da un conduttore avvolto a elica, con aspetto simile a quello di una molla a spirale). Quando collocò i due solenoidi uno accanto all'altro e vi fece passare una corrente che fluiva nella stessa direzione (*si veda l'illustrazione a pagina 93*), Ampère si attendeva che i solenoidi si respingessero fra loro, come due magneti situati l'uno accanto all'altro con i poli nord puntati nella stessa direzione. Essi invece si attrassero reciprocamente.

Ampère tentò allora una seconda via. Avvolse il suo conduttore elettrico di rame in spirali piane, con un'estremità che usciva dal centro della spirale e l'altra che usciva dalla periferia. Questa volta, quando fece passare corrente e tenne le spirali una vicina all'altra, esse si comportarono effettivamente come magneti, respingendosi. Il 25 settembre egli diede ai suoi colleghi una dimostrazione di vari effetti di attrazione e repulsione.

Perché i solenoidi si erano comportati in modo diverso? La risposta venne da un'osservazione casuale. I conduttori che collegavano le due spirali erano connessi alla stessa batteria, cosicché i tratti di conduttore connessi allo stesso polo della batteria passavano l'uno accanto all'altro. Quando Ampère fece passare corrente, osservò che i conduttori percorsi da corrente nella stessa direzione si attraevano fra loro, anche se non erano avvolti circolarmente. Questa interazione fra conduttori elettrici rettilinei, che non era mai stata osservata prima, diede ad Ampère la soluzione al problema dei solenoidi. Poiché aveva usato filo di rame non isolato, egli aveva avvolto i solenoidi in modo che le singole spire in ognuno di essi fossero relativamente distanziate fra loro. Il moto circolare della corrente era stato perciò insignificante rispetto al percorso longitudinale da un estremo del solenoide all'altro. Quindi i solenoidi erano risultati molto simili a due conduttori rettilinei percorsi da corrente nella stessa direzione, e questo spiegava perché si attraessero a vicenda.

Per verificare sperimentalmente questa conclusione, Ampère preparò un solenoide avvolto attorno a un tubo di vetro e fece ripassare l'estremità del filo al-

l'interno del tubo. Egli si attendeva che la corrente che passava all'interno del tubo cancellasse l'effetto longitudinale della corrente che fluiva da un estremo del solenoide all'altro, mettendo in evidenza gli effetti della componente circolare della corrente. I solenoidi costruiti in questo modo si comportarono esattamente come magneti permanenti, confermando l'ipotesi di Ampère.

Questi risultati sollevarono un difficile interrogativo: dove sono le correnti elettriche in un magnete permanente? In realtà esistono solo due possibilità: o le correnti si muovono in cerchio attorno all'asse di tutto il magnete oppure fluiscono in cerchi molto più piccoli attorno a ciascuna delle particelle da cui il magnete è composto. Su suggerimento di Fresnel, Ampère suppose che le correnti fluiscano attorno a singole molecole del magnete.

Un'ipotesi del genere richiedeva una conferma sperimentale; Ampère fondò la sua verifica su un esperimento eseguito da Arago. Questi aveva realizzato un solenoide di filo di rame e vi aveva sospeso al centro un ago di ferro. Quando aveva collegato le due estremità del solenoide ai poli di una batteria, la corrente elettrica circolare risultante aveva magnetizzato l'ago. Perciò, se la teoria di Ampère era corretta, la corrente circolare nel solenoide aveva generato correnti circolari all'interno dell'ago. Ma le correnti fluivano attorno all'asse dell'ago o attorno alle singole molecole al suo interno?

Ampère tentò di trovare la risposta a questa domanda piegando ad anello un sottile nastro di rame e sospendendolo all'interno di una spira circolare di filo di rame isolato. Il diametro dell'anello era leggermente più piccolo di quello della spira; l'anello e la spira erano allineati in modo che essi fossero concentrici e i loro assi paralleli. Se nell'esperimento di Arago la corrente che percorreva un solenoide aveva generato correnti circolari attorno all'asse dell'ago, una corrente simile doveva provocare un flusso circolare di correnti nel nastro di rame, facendolo operare temporaneamente come un magnete. Ampère verificò la presenza di tale corrente tenendo un magnete a barra dentro l'anello mentre la spira era percorsa da corrente. Se l'anello fosse stato magnetizzato, sarebbe stato deviato dal magnete a barra, mentre, come Ampère aveva previsto, non ci fu alcuna deviazione. Ampère citò pubblicamente questo esperimento come una forte prova a sostegno della sua ipotesi delle correnti molecolari.

Quasi contemporaneamente, però, la sua intera teoria concernente le cause del magnetismo permanente fu pubblicamente contraddetta. Verso la fine del 1821 fu pubblicata in Inghilterra, e tradotta immediatamente in Francia, una storia dell'elettromagnetismo anonima, scritta, come oggi sappiamo, dal fisico inglese Michael Faraday. Il libro descriveva un paio di esperimenti progettati per confutare l'ipotesi basilare di Ampère che il magnetismo permanente fosse semplicemente il risultato di correnti elettriche circolari. Secondo Faraday i risultati sperimentali dimostravano che il magnetismo di un magnete permanente è chiaramente diverso dal magnetismo di un solenoide percorso da corrente.

Nel primo esperimento veniva realizzato un solenoide avvolgendo del filo di rame isolato attorno a un grosso tubo cavo di vetro. Il tubo veniva immerso per metà in un recipiente contenente acqua, con l'asse maggiore parallelo alla superficie dell'acqua. Poi un lungo ago magnetizzato veniva fissato su un sughero galleggiante in prossimità del tubo e attraverso il solenoide veniva fatta passare una corrente elettrica. Se il solenoide fosse stato esattamente analogo a un magnete, sosteneva Faraday, il sughero avrebbe dovuto spostarsi sino al tubo e poi fermarsi, poiché un polo dell'ago (per esempio il polo nord) sarebbe stato allora il più vicino possibile al polo opposto del solenoide-magnete (il polo sud). Invece il sughero, dopo essere giunto a un'estremità del tubo, continuò a muoversi al suo interno, finché i due poli dell'ago andarono a fermarsi direttamente sotto i poli uguali (e non sotto quelli opposti) del solenoide-magnete.

Il secondo esperimento esaminò l'effetto nel caso di un vero magnete permanente. Un foglio di acciaio fu arrotolato a formare un tubo, magnetizzato e sospeso in modo da essere immerso per metà nell'acqua. Questa volta il sughero si mosse verso il tubo, ma non entrò in esso, fermandosi esattamente sotto il polo sud del magnete; secondo Faraday, questa era una conferma del fatto che un solenoide percorso da corrente non è esattamente uguale a un magnete permanente e quindi che il magnetismo permanente non è il prodotto di correnti elettriche circolari.

Ampère trovò una via d'uscita: se in un magnete permanente le correnti circolano attorno alle singole molecole, l'interno di un tubo d'acciaio *dovrebbe* essere qualitativamente diverso da quello di un solenoide. Nel solenoide l'ago magnetizzato è racchiuso all'interno della corrente circolante, mentre nel tubo rimane all'esterno di tutte le correnti molecolari. Era perciò lecito aspettarsi che nei due casi l'ago agisse diversamente. Avendo fornito questa spiegazione (essenzialmente corretta), Ampère era ora pubblicamente impegnato a difendere il concetto di correnti molecolari.

La forza dell'impegno di Ampère risultò evidente nell'estate del 1822, quando egli ripeté l'esperimento nel quale utilizzava un cerchio di nastro di rame e una spira di rame. Questa volta, però, Ampère verificò il comportamento del nastro con un potente magnete a ferro di cavallo anziché con un più debole magnete a barra, e il nastro venne effettivamente deviato. La reazione di Ampère a questo risultato fu abbastanza sconcertante. Esso sembrava contraddire la sua teoria delle correnti molecolari, ma egli non reagì sottoponendo la teoria a ulteriori esperimenti. Menzionò solo di sfuggita questo esperimento nella sua relazione all'Académie des Sciences nel mese di settembre, con la sorprendente osservazione che l'effetto non aveva alcun significato teorico. In realtà Ampère aveva osservato senza rendersene conto il fenomeno dell'induzione elettromagnetica tra due correnti elettriche, ma fu solo nel 1832, dopo che Faraday aveva scoperto e studiato l'induzione elettromagnetica, che Ampère si rese conto di essere stato sul punto di compiere una scoperta importante e di essersela lasciata sfuggire.

I concetti fondamentali dell'elettrodinamica di Ampère non mutarono mai dopo il 1822. Quel che mutò fu la capacità di compiere valutazioni quantitative con la teoria. Una volta formulata la teoria del magnetismo permanente, il compito era quello di determinare sperimentalmente quanto siano intense le forze elettromagnetiche. Dopo avere deciso che l'interazione fondamentale in elettrodinamica è la forza che agisce fra due conduttori percorsi da corrente, Ampère si dedicò al difficile compito di misurarla. Ancora una volta gli fu d'aiuto la sua capacità di formulare ipotesi. A differenza delle forze gravitazionali, che possono essere trattate matematicamente come forze agenti fra semplici punti geometrici, le forze dovute a una corrente che percorre un conduttore non possono sempre essere considerate alla stregua di fenomeni locali semplici.

L'idea di Ampère fu quella di considerare segmenti infinitamente piccoli di conduttori elettrici percorsi da corrente e di supporre che le forze agenti fra tali segmenti variassero con l'inverso del quadrato della distanza fra i segmenti stessi. Egli trovò poi la forza totale fra i due conduttori integrando, ossia sommando matematicamente, tutte le unità infinitesime di forza e tenendo conto delle direzioni in cui le forze avrebbero potuto agire in funzione del modo in cui erano piegati i conduttori. Ampère considerò dapprima il caso speciale in cui i due conduttori giacciono nello stesso piano. Poi generalizzò il suo risultato per tener conto di conduttori giacenti in piani distinti, quale che fosse la loro inclinazione reciproca. Questa generalizzazione gli permise di considerare fili ripiegati in qualsiasi modo nello spazio tridimensionale. Il suo risultato finale fu la famosa e concisa formula con la quale si può calcolare la forza elettrodinamica che agisce fra due conduttori purché si conosca l'intensità delle correnti e la disposizione geometrica dei conduttori. Nel 1826 Ampère rielaborò le sue pubblicazioni precedenti riunendole nella sua opera più completa: *Mémoire sur la théorie mathématique des phénomènes électrodynamiques uniquement déduite de l'expérience* (Memoria sulla teoria matematica dei fenomeni elettrodinamici dedotta unicamente dall'esperienza).

Dopo il 1827 la salute di Ampère cominciò a declinare rapidamente. Egli abbandonò la ricerca scientifica creativa per dedicarsi alla sua ultima opera sulla filosofia della scienza. Qui egli riscoprì una parte dell'ispirazione della sua giovinezza e fu conquistato dalla dottrina dell'«armonia prestabilita» di Gottfried Wilhelm von Leibniz, secondo la quale la mente dell'uomo sarebbe una copia, per quanto imperfetta, della mente di Dio. Poiché, secondo Leibniz, il processo di ragionamento dell'uomo è un'immagine del processo di ragionamento di Dio e poiché è stata la ragione di Dio a creare l'universo, la mente umana dovrebbe essere in grado di comprendere l'universo attraverso un processo di ragione pura: in altri termini, dovrebbe esserci un'armonia preesistente fra le leggi dell'universo e i poteri di ragionamento dell'uomo.

Ampère suppose che la corrispondenza fra la ragione di Dio, la ragione umana e la razionalità intrinseca dell'universo dovrebbe rendere possibile l'uso di quello che è essenzialmente un processo tassonomico per trovare la verità suprema. A suo parere, se si potessero delineare tutte le scienze accessibili alla mente umana si avrebbe la chiave fondamentale per ogni possibile verità, essendo la mente umana strutturata in un modo che corrisponde direttamente alla struttura dell'universo. Resterebbero solo da inserire i contenuti del diagramma tassonomico cosmico.

Ampère produsse molti di questi diagrammi - che nel suo intendimento dovevano essere strumenti della ricerca fondamentale - prima di morire, probabilmente di polmonite, nel 1836. Egli terminò così la sua vita intellettuale quasi come l'aveva cominciata: come un enciclopedista fermamente convinto dell'unità di tutto il sapere, poiché ogni forma di conoscenza non è altro che il riflesso dell'unità della mente divina.

BIBLIOGRAFIA

WILLIAMS L. PEARCE, *Ampère* in «Dictionary of Scientific Biography» a cura di Charles Coulston Gillispie, Charles Scribner's Sons, 1970.

BLONDEL CHRISTINE, *A-M. Ampère et la création de l'électrodynamique (1820-1827)*, Comité des travaux historiques et scientifiques, 1982.

WILLIAMS L. PEARCE, *What Were Ampère's Earliest Discoveries in Electrodynamics?* in «Isis», 74, n. 274, dicembre 1983.

(RI)CREAZIONI AL CALCOLATORE

di A. K. Dewdney

Variazioni sul tema dei rompicapo di persone

«L'analisi dei caratteri sembra essere quanto di più dilettevole ci sia per l'uomo».

Isaac Bashevis Singer,
«New York Times Magazine»,
26 novembre 1978

Un giorno, tre filosofi dell'antica Grecia stavano passeggiando nei dintorni di Atene sotto il sole di mezzogiorno. Sedutisi all'ombra di un ulivo, stapparono una fiaschetta di vino e iniziarono una tranquilla discussione sul problema ontologico fondamentale: perché le cose esistono? La discussione si fece accesa, poi divenne sempre più confusa e sconnessa, finché, poco dopo, tutti e tre i filosofi finirono con l'addormentarsi all'ombra dell'albero.

Nel pomeriggio, un giovane ateniese molto dispettoso, che si era messo a spiare i tre filosofi addormentati, schizzò sulle loro fronti, con delicatezza, della vernice bianca. Poco prima del tramonto, una civetta che si era fatta il nido nell'albero zampettò su un ramo proprio sopra i tre uomini, lanciò uno stridio poderoso e svolazzò via rumorosamente. Lo stridio della civetta svegliò i filosofi; ciascuno dei tre pensò immediatamente che l'uccello fosse responsabile della «decorazione» sulla fronte dei suoi due colleghi e tutti e tre scoppiarono simultaneamente a ridere.

Lo spettacolo era senza dubbio divertente, e ci vollero circa cinque secondi prima che uno dei tre smettesse improvvisamente di ridere. Perché? Il rompicapo può essere risolto spiegando il processo mentale del filosofo che smise di ridere. Nel far questo, si deve prendere in considerazione ciò che egli deve aver supposto che i suoi compagni stessero pensando.

In molti rompicapo entrano in gioco persone solo per fornire al potenziale solutore un contesto umano in cui possa trovarsi più a suo agio. Ma la soluzione dei problemi che chiamo rompicapo di persone, come quello dei tre filosofi, richiede che si pensi a ciò che ciascuna persona del rompicapo sta pensando che le altre persone del rompicapo stiano pensando. Questo ragionamento sul ragionamento non è solo un argomento divertente per i rompicapo, ma è anche un tema di studio fondamentale per gli esperti di scienza dei calcolatori che vogliano creare programmi capaci di imitare il modo in cui un essere umano ragiona in situazioni complesse.

Il tema di fondo di questo articolo non si può considerare completo senza la soluzione del rompicapo dei tre filosofi. Diamo un nome a questi filosofi: Pitagora, Platone e Aristotele. Pitagora, il più vecchio e il più saggio, è quello che smise di ridere. Ecco perché: Pitagora vedeva Aristotele sghignazzare e si rese conto che Aristotele non supponeva assolutamente di avere la fronte macchiata di bianco. Se lui (Pitagora) aveva la fronte pulita, allora Aristotele stava evidentemente ridendo di Platone. Ma allora Aristotele di chi pensava che Platone stesse ridendo? «Per Atena!», deve avere esclamato Pitagora tra sé e sé, «io non dovrei ridere». La situazione è illustrata sotto forma di fumetto nella pagina a fronte.

Questa è la risposta alla forma più o meno tradizionale del rompicapo, un rompicapo che può essere esteso chiedendo perché pochi secondi dopo anche Platone smette di ridere. Appena Pitagora smette di ridere, Platone non può più seguire il suo stesso ragionamento. Sembrerebbe, in effetti, che ora Platone veda rinforzata la propria convinzione di avere la fronte pulita. «Evidentemente Pitagora ha visto che la mia fronte è pulita e si è reso conto che Aristotele stava ridendo di lui», potrebbe aver pensato Platone. Ma se ci ragiona ancora un poco, deduce di essere anche lui imbrattato. Affido al lettore il compito di calarsi nei panni di Platone per capire perché.

Consideriamo una variante della storia dei tre filosofi. Un sultano voleva scegliere il più saggio fra tre visir per la carica di gran visir. Li portò in una stanza buia e mise sulla testa di ciascuno di essi un copricapo bianco. Li riportò poi nella sala del trono e disse loro: «Ciascuno di voi ha ricevuto un copricapo che è o bianco o nero, e almeno uno dei copricapi è bianco. Il primo di voi che mi dirà il colore del proprio copricapo avrà la carica di gran visir».

Il rompicapo dei tre visir è identico nella sostanza a quello dei tre filosofi, anche se non c'è alcuna risata che serva a indicare ciò che ciascun visir vede. Curiosamente, il fatto che tutti sappiano che almeno un copricapo è bianco, congiuntamente al silenzio dell'indecisione, produce un effetto equivalente.

Chiamiamo al-Khwarizmi, ibn Khaldun e ibn Sina i tre visir candidati. Ecco il ragionamento di al-Khwarizmi, che gli frutta la carica: «Hmm, vediamo: può essere nero il mio copricapo? Se lo fosse, che cosa penserebbero gli altri due visir? Supponiamo che anche ibn Khaldun pensi che il suo copricapo sia nero. In quel caso si renderebbe conto che ibn Sina vedrebbe due copricapi neri e dedurrebbe immediatamente che il suo è bianco.
to "Il mio copricapo è bianco!", quindi ibn Khaldun sa che ibn Sina vede almeno un copricapo bianco. Ma se il mio fosse davvero nero, allora ibn Khaldun saprebbe che il suo è il copricapo bianco che ibn Sina vede, e lo direbbe. Invece ibn Khaldun non lo ha fatto. Per la barba del Profeta, il mio copricapo non può essere nero!»

Come ha rilevato un mio collega, il rompicapo dei tre visir ha un metalivello: appena ai visir viene comunicata la natura della competizione, ciascuno di essi può dedurre che l'unica prova equa delle loro abilità cognitive richiede in realtà che tutti e tre i copricapi siano bianchi.

Il rompicapo dei tre filosofi e quello dei tre visir si possono facilmente riscrivere in un numero infinito di modi. In una nota variante ci sono tre dame dell'aristocrazia che, verso la fine del secolo scorso, viaggiano in treno nella campagna inglese. La giornata è molto calda ed esse hanno aperto il finestrino dello scompartimento per far entrare un po' di aria fresca. Il treno, sbuffando un denso fumo nero, entra in una lunga galleria. Quando riemerge alla luce del sole, le tre dame scoppiano simultaneamente a ridere alla vista delle facce coperte di fuligine delle loro compagne di viaggio - finché una smette di ridere.

Ecco una seconda variante: tre uomini d'affari mangiano una *quiche* di spinaci in un ristorante di Manhattan. A tutti e tre rimangono degli spinaci tra i denti...

Il fatto che un rompicapo di persone può assumere molte diverse forme equivalenti ci porta al concetto di trasformazioni di un rompicapo. Come si fa, per così dire, a trasformare filosofi in visir? È utile identificare gli elementi di un rompicapo che corrispondono a elementi dell'altro:

1. giovane dispettoso → sultano
2. filosofi → visir
3. fronte → testa
4. segno bianco → copricapo bianco
5. ridere → rimanere in silenzio

Oltre a questi elementi, la trasformazione deve anche dare una definizione operativa della caratteristica peculiare di un filosofo o di un visir. Un filosofo è qualcuno che, finché non analizza la situazione, continua a ridere se vede almeno una fronte imbrattata, mentre non ride in caso contrario. Un visir è qualcuno che, finché non analizza la situazione, non dice niente se vede almeno un copricapo bianco. In caso contrario dice: «Sulla mia testa c'è un copricapo bianco».

I lettori sono invitati a riempire gli spazi lasciati vuoti nel generico rompicapo di persone che segue. Riempiendo gli spazi vuoti numerati con le corrispondenti parole della colonna di sinistra il risultato sarà il rompicapo dei tre filosofi. Sostituendo invece agli spazi vuoti le parole della colonna di destra si otterrà il rompicapo dei tre visir. (Ci si può naturalmente inventare un altro insieme di parole che, inseriti negli spazi vuoti, dia senso alla storia.) I dettagli di tempo e luogo, unicamente intesi a vivacizzare il racconto, sono stati eliminati per semplificare la dimostrazione.

Un giorno un (1).......... mise un (4).......... sulla (3).......... di ciascuno di tre (2).......... senza che essi lo sapessero. Non sapendo che sulla propria (3)..........
dei (2).......... continuò a (5).......... .
Improvvisamente, il più saggio dei tre (2)..........
esclamò: «Sulla mia (3).
(4)..........!» Come ha fatto a capirlo?

La variazione sul tema dei tre filosofi che presenterò ora aveva a che fare, nella sua versione originale, con mogli infedeli. Dato che comporta degli stereotipi sessuali, la modificherò operando una trasformazione molto semplice. In ogni caso, il rompicapo è interessante perché costituisce una generalizzazione del problema dei tre filosofi.

La tirannica regina delle Amazzoni annuncia un giorno che nel suo regno almeno un marito si è macchiato di infedeltà. Emana dunque un proclama: «Ordino che se qualcuna di voi scopre che suo marito le è stato infedele, lo uccida allo scoccare della mezzanotte del giorno in cui ha accertato l'infedeltà». Nel regno delle Amazzoni, le informazioni erano condivise liberamente, ma non troppo: ogni moglie era a conoscenza delle infedeltà di tutti i mariti tranne che del proprio. Inoltre, la notizia di un'esecuzione si diffondeva in tutto il reame entro un giorno. Si dava il caso che ci fossero esattamente 40 mariti infedeli. Qualcuno di essi fu giustiziato? E in caso affermativo, quando?

I lettori avranno notato che l'annuncio della regina dice che almeno un marito è stato infedele. Se esattamente un marito avesse tradito sua moglie, questa l'avrebbe saputo subito: infatti, se il traditore fosse stato il marito di qualcun'altra, e non il suo, ne avrebbe avuto notizia. Quindi, alla mezzanotte del giorno stesso in cui fu emanato il proclama, il marito infedele sarebbe stato ucciso da sua moglie.

Se ci fossero stati esattamente due mariti infedeli, le rispettive mogli li avrebbero fatti fuori alla mezzanotte del secondo giorno. La notizia che non c'era stata alcuna esecuzione alla mezzanotte del primo giorno, infatti, avrebbe confermato il fatto che c'erano due mariti infedeli.
fraghi sarebbero state a conoscenza di un solo traditore nel regno (anche se tutte le altre Amazzoni si sarebbero rese conto che ce n'erano almeno due), avrebbero immediatamente capito che il loro sposo doveva essere il secondo traditore.

I lettori avranno ormai afferrato il bandolo del ragionamento. Nessuna esecuzione alla mezzanotte dell'n-esimo giorno significa che almeno $n + 1$ mariti sono infedeli.
giorno, sarebbe stato di dominio pubblico il fatto che almeno 40 mariti erano infedeli. Questo non avrebbe costituito alcuna sorpresa per le Amazzoni sposate a mariti fedeli, in quanto sarebbero state a conoscenza di 40 mariti fedifraghi. Solo una moglie con un marito infedele avrebbe saputo di 39 traditori, e questo avrebbe voluto dire che il suo era il quarantesimo marito infedele. Queste mogli, in ossequio all'ordine della loro sovrana, avrebbero dunque convocato i loro mariti per un tête-a-tête di mezzanotte il quarantesimo giorno dal proclama della regina.

Il rompicapo delle Amazzoni è realmente una variante di quello dei tre filosofi? Si può rispondere alla domanda chiedendosi cosa avverrebbe se ci fossero solo tre Amazzoni, ciascuna sposata a un marito infedele. In quel caso, allo scadere del terzo giorno ciascuna delle Amazzoni sarebbe arrivata alla conclusione esatta. In questa forma, il rompicapo è direttamente assimilabile a quello dei tre visir (che, come ho detto, è una
Eppure ibn Sina non ha esclama-
variante del rompicapo dei tre filosofi).

Per verificarlo, supponiamo che il sultano abbia detto ai candidati alla carica di gran visir: «Vi chiederò diverse volte in successione se sapete qual è il colore del vostro copricapo. Rispondete solo se lo sapete, altrimenti state zitti». In quel caso, la prima volta che il sultano avesse posto la domanda, tutti e tre i visir sa-

Un filosofo pensa a un livello più profondo dei suoi colleghi

rebbero rimasti in silenzio. Lo stesso sarebbe avvenuto la seconda volta. Alla terza volta, tutti e tre i candidati avrebbero dato la risposta.

Il rompicapo delle Amazzoni riguarda 40 mariti infedeli, non tre. Si può generalizzare il rompicapo dei tre filosofi a un gruppo di 40 filosofi? Immaginiamo per il momento che i filosofi addormentati sotto l'albero siano solo quattro. Al loro risveglio, cominciano tutti a ridere e il quarto filosofo (in realtà un dio sotto false spoglie) ragiona in questo modo: «Hmm. È consono alla mia dignità divina presupporre di avere la fronte pulita e quindi posso ridere dei tre mortali insudiciati. Ma perché uno di loro non si rende conto del proprio stato e non smette di ridere?» (L'abitante dell'Olimpo ricapitola ora mentalmente il ragionamento di Pitagora.) «Povero me, forse ne so il motivo.»

Se questo processo mentale può portare un quarto filosofo a rendersi conto di non essere stato risparmiato, può altrettanto facilmente spiegare perché un quinto, un sesto e anche un quarantesimo filosofo possano arrivare alla stessa conclusione. In una puntata della sua rubrica di «Giochi matematici», Martin Gardner ha elaborato in modo analogo una generalizzazione del rompicapo dei tre visir (si veda «Le Scienze» n. 109, settembre 1977). Come egli osservava acutamente, sorgono però delle difficoltà. «Questa generalizzazione solitamente suscita discussioni perché il problema richiede tante imprecise assunzioni sul grado di intelligenza e sulla durata dei lassi di tempo da renderlo irreale.»

[In quello stesso fascicolo, Martin Gardner procedeva presentando una diversa forma del problema dei cappelli colorati, al fine di evitare ogni confusione: «Ci sono tre uomini *A*, *B*, e *C* e cinque cappelli. Tre cappelli sono rossi e due neri. Si suppone che ognuno dei tre sia onesto e "razionale", cioè che possa fare rapidamente qualsiasi deduzione valida indipendentemente da quanto essa sia complicata... . Gli uomini chiudono gli occhi e un "arbitro" mette a ciascuno un cappello rosso. Gli altri due cappelli vengono nascosti... a ciascuno viene chiesto, nell'ordine: "Sai il colore del tuo cappello?"

«*A* risponde sinceramente di no, *B* pure di no, *C* invece dice: "Sì, il mio cappello è rosso." Come fa a saperlo?» Gardner nota anche che questo problema ha un aspetto sorprendente: *C* potrebbe rispondere affermativamente anche se fosse cieco! Se il problema dei cappelli colorati vi incuriosisce, vale la pena di rileggere quella puntata della rubrica di Gardner, in cui vengono presentate anche altre forme del problema, non prese in considerazione qui da Dewdney. C'è poi una ripresa del problema in «Giochi matematici» nel numero successivo di «Le Scienze», il 110 dell'ottobre 1977. *n.d.r.*]

A dire il vero, tutti i rompicapo hanno in sé un certo grado di inverosimiglianza. Quanto può essere verosimile, anche in un ristorante alla moda, che tre persone che mangiano una *quiche* di spinaci sorseggiando dello Chardonnay ben fresco finiscano tutte e tre con pezzetti di spinaci tra i denti nello stesso momento? E anche se questo avvenisse, che probabilità c'è che uno di loro non veda dall'occhiata di un altro che si sta ridendo anche di lui?

Eppure, per i ricercatori che si sforzano di elaborare nuove forme di intelligenza artificiale basate su quella che va sotto il nome di programmazione logica, i rompicapo di persone sono questioni serie. Tra gli altri, John L. McCarthy dell'Università della California a Berkeley ha messo alla prova la forza deduttiva di sistemi logici mettendoli al lavoro su rompicapo di persone come quello dei tre visir. I sistemi di programmazione logica sfruttano un tipo di ragionamento simbolico noto come calcolo predicativo per derivare automaticamente varie deduzioni da presupposti dati. Per imitare la capacità umana di trarre deduzioni in varie situazioni sociali, i sistemi di programmazione logica devono essere in grado di costruire modelli del ragionamento sul ragionamento.

Questi ricercatori potrebbero anche prendere in considerazione il tipo più realistico di rompicapo di persone che si presenta nella vita quotidiana. Anche se questi rompicapo non hanno profondità logica e richiedono numerosi presupposti per essere risolti, nondimeno rivelano che noi dedichiamo una parte significativa della nostra capacità mentale a pensare a ciò che altre persone possono pensare che noi stiamo pensando.

Erving Goffman, che lavora anch'egli a Berkeley, è un pioniere di quel campo della psicologia chiamato analisi transazionale, la quale afferma che ogni persona cerca continuamente di sapere quello che gli altri pensano di lui, o di lei, e di manipolare quel pensiero, un po' alla maniera degli attori. Goffman ha documentato questo aspetto della nostra vita mentale con un elenco di comportamenti peculiarmente umani che riconosciamo come esperienze valide.

Quelli che seguono sono due esempi del tipo descritto da Goffman tratti dalla mia vita reale; si dà il caso che io abbia raccolto per anni esempi di questo genere, che mi sembrano adatti a mettere in luce la capacità umana di pensare a ciò che gli altri stanno pensando, capacità di cui nella maggior parte dei casi, io credo, siamo solo parzialmente consci.

Una volta vidi un poliziotto scendere in strada da un edificio di aspetto un po' equivoco. Forse si trattava di un salone di massaggi. (A giudicare dal quartiere in cui avvenne il fatto, la cosa mi sembra verosimile.) Appena uscito in strada, il poliziotto si trovò inaspettatamente davanti un collega, al quale si rivolse con un'espressione stranissima: non appariva imbarazzato, ma sembrava quasi che volesse far finta di esserlo. La cosa mi sembrò strana sul momento, ma in seguito mi risultò comprensibile. Inizialmente, la reazione del primo poliziotto doveva essere stata di vero imbarazzo. Dopo tutto, il suo collega avrebbe avuto tutte le ragioni di sospettare che la visita fosse stata, diciamo così, non ufficiale. Ma esprimere questo imbarazzo non avrebbe fatto che confermare i sospetti. Se invece il secondo poliziotto avesse ricevuto l'impressione che il primo poliziotto stava solo facendo finta di essere imbarazzato, avrebbe forse preso la cosa come uno scherzo.

I giochi mettono a volte in luce l'aspetto peggiore delle persone. Due colleghi piuttosto competitivi avevano appena finito una partita di poker. Il perdente era visibilmente alterato ma cercava di controllarsi. Alla fine sbottò: «Avevi in mente di ....mi e io avevo in mente di ....ti. Semplicemente ci sei riuscito tu!» (Gli spazi vuoti possono essere riempiti con un'adeguata parola volgare di significato analogo, in questo contesto, a danneggiare.)

La parola significativa, dal punto di vista operativo, è in questo caso *semplicemente*. A prima vista sembra implicare che nella vittoria dell'avversario il ruolo principale è stato giocato dalla fortuna. A un esame più approfondito, però, risulta chiaro che con quell'espressione il perdente vuole che il vincitore pensi che lui (il perdente) pensa che l'aspetto più significativo della loro partita sia stata l'identità delle loro motivazioni e non la disparità che la sua sconfitta implicava.

Mi chiedo se qualche lettore possa inviarmi esempi tratti dalla propria esperienza. Particolarmente interessanti sarebbero esempi che abbiano un po' più di attinenza, nel loro contenuto, con i tradizionali rompicapo di persone.

A proposito di rompicapo, è uscito uno splendido libro sull'argomento, *The Puzzling Adventures of Dr. Ecco*. L'autore è Dennis Shasha, uno studioso di informatica del Courant Institute of Mathematical Sciences della New York University. Il libro descrive le avventure del misterioso dottor Ecco, un eccentrico di incredibile intelligenza che si guadagna da vivere risolvendo rompicapo. Oltre a un paio di rompicapo di persone, ci sono rompicapo che riguardano elezioni, strade multiple, spie, circuiti che controllano circuiti e molte altre cose. C'è anche un concorso in cui i partecipanti sono sfidati a risolvere 10 rompicapo di decodificazione. Chi riesce a risolverli vince una «maglietta onnieuristica» (simbolo di suprema abilità intellettuale) e il diritto di partecipare all'estrazione di un set di scacchi in onice lavorato a mano. Il concorso si chiude l'uno aprile 1989. Il nome del vincitore del primo premio apparirà in questa rubrica in settembre.

Per terminare l'esame dei rompicapo di persone ne riporterò uno tratto dal libro di Shasha. Due generali, i cui rispettivi eserciti sono schierati sui fianchi opposti di un rilievo montuoso, vogliono coordinare il loro attacco perché se uno dei due eserciti attaccasse da solo potrebbe essere distrutto. Purtroppo i due generali possono comunicare solo con piccioni viaggiatori che volano da un campo all'altro superando il rilievo montuoso (*si veda l'illustrazione della pagina a fronte*).

Avrà mai fine la serie di conferme tra i generali?

Il primo generale invia questo messaggio al secondo generale: «Attacco alle ore 08.00. Confermare il messaggio ricevuto: in caso contrario non attaccherò». Il secondo generale non ha obiezioni ad attaccare a quell'ora e vuole inviare un piccione con la conferma, ma improvvisamente si rende conto che il primo generale non attaccherà se non riceve la conferma. Non avendo la garanzia che il piccione viaggiatore riuscirà realmente a consegnare la sua conferma al primo generale, decide che anche lui non attaccherà prima di sapere che il primo generale sa che l'attacco è stato confermato. Il secondo generale manda allora un piccione con un messaggio per il primo generale. Avrà mai fine questa serie apparentemente infinita di messaggi di conferma? Forse la risposta dipende dal particolare messaggio che uno dei due generali invia a un certo punto.

Ringrazio per l'aiuto fornitomi nella preparazione di questa rubrica due dei miei colleghi dell'Università dell'Ontario occidentale, Andy L. Szilard e Areski Nait Abdallah.

Proclamerò ora i vincitori della sfida per la griglia di numeri primi proposta da Gordon Lee e presentata in questa rubrica nel settembre scorso. Una griglia di numeri primi è una griglia di 6 per 6 caselle, ciascuna contenente un'unica cifra. L'obiettivo è quello di scegliere i numeri primi e la loro posizione nella griglia in modo tale da ottenere quanti più numeri primi possibile leggendo cifre consecutive lungo qualsiasi linea (orizzontale, verticale o diagonale). Lee, che ha organizzato una gara analoga in Inghilterra l'anno scorso, mi ha riferito che la griglia vincitrice conteneva 170 numeri primi. I lettori di questa rubrica potevano fare di meglio? Io non ne ho mai dubitato!

Ho ricevuto molte più risposte di quanto pensassi, sia da geniacci tecnologici che hanno utilizzato supercalcolatori, sia da gente più umile che si è data da fare con carta e matita. Ho spedito tutto a Lee per il giudizio ed ecco i risultati in ordine inverso:

Il più prolifico: Larry J. Padden di Oklahoma City (Oklahoma), che ha prodotto 147 diverse griglie di numeri primi che contengono da 170 a 173 numeri primi.

Per le sestuple: David Mckenzie e Frank Endres di Austin (Texas), che hanno scoperto una griglia con solo 106 numeri primi, ma in cui tutti i 28 possibili numeri di sei cifre sono primi.

Terzo classificato: James I. Waldby di Robinson (Illinois), che ha totalizzato 185 numeri primi.

Secondo classificato: Mckenzie e Endres dividono questo onore con Stephen C. Root di Westboro (Massachusetts). Sono riusciti a ottenere griglie con 186 numeri primi.

Primo classificato: Ancora Root. La sua griglia vincente, che è riportata qui sotto, ha un totale di 188 numeri primi:

3	1	7	3	3	3
9	9	5	6	3	9
1	1	8	1	4	2
1	3	6	3	7	3
3	4	9	1	9	9
3	7	9	3	7	9

Molti lettori hanno individuato il percorso più breve per attraversare il crudele labirinto tridimensionale descritto nella rubrica di novembre. Era lungo 39 caselle (comprese quelle di entrata e di uscita). Nel labirinto ci sono molti altri percorsi che arrivano all'uscita, ma tutti passano per il punto in cui, annidato tra le tenebre, il feroce Minotauro è in attesa dei fanciulli e delle fanciulle ateniesi obbligati a entrare nel labirinto. Alcuni lettori hanno scoperto la collocazione segreta del Minotauro tracciando laboriosamente una mappa per poi individuare il passaggio obbligato; altri, invece, hanno sospettato che l'alto grado di simmetria presente nel disegno del labirinto potesse essere un'indicazione per stabilire il covo del Minotauro. Così era.

C'è solo un punto in cui il labirinto tridimensionale è asimmetrico. I lettori che tornino a rivedere il labirinto scopriranno una parete «mancante» tra le caselle (5,3) e (5,4) del secondo livello. (Le coordinate stanno a indicare rispettivamente di quante caselle ci si debba muovere verso destra dal lato sinistro del labirinto e di quante caselle ci si debba muovere verso il basso dal lato superiore.) Perché il passaggio obbligato deve trovarsi in quel punto di asimmetria?

I lettori che hanno inviato per primi la soluzione esatta al problema del Minotauro sono Michael Amling di Glen Ellyn (Illinois); Lawrence Leinweber di Cleveland Heights (Ohio); Thomas R. Lunsford, Jr., di Hinesville (Georgia); Donald E. G. Maln di Rochester (Michigan); Jim Newton di Middleton (Wisconsin) e Ken Silber di New York.

BIBLIOGRAFIA

GOFFMAN ERVING, *The Presentation of Self in Everyday Life*. Doubleday & Company, Inc., 1959.

SHASHA DENNIS, *The Puzzling Adventures of Dr. Ecco*. W. H. Freeman and Company, 1988.